INDUS JOURNAL OF SCIENCE
https://induspublishers.com/IJS
Volume 1, Issue 2 (2023)

Original Article

DATA MINING IN HEALTHCARE: APPLYING DATA MINING AND MACHINE LEARNING TECHNIQUES TO ANALYZE LARGE HEALTHCARE DATASETS, SUCH AS ELECTRONIC HEALTH RECORDS, FOR IMPROVED DIAGNOSIS,

Shafi Ullah[a], Muhammad Nadeem[a*], Muhammad Asif[a]

[a] Regional Blood Centre Dera Ismail Khan, Khyber Pakhtunkhwa, Pakistan

ARTICLE INFO

Key Words:

* Data mining
* Electronic health records
* Healthcare practices
* Python programming
* Software development

***Corresponding Author:**
Muhammad Nadeem
nadeemmasood008@gmail.com

ABSTRACT

Background: The provided text discusses the application of data mining in healthcare, specifically in analyzing large healthcare datasets, such as electronic health records (EHRs), for improved diagnosis, treatment, and patient outcomes.

Objectives: To apply data mining and machine learning techniques to analyze large healthcare datasets, such as electronic health records, for improved diagnosis, treatment, and patient outcomes.

Methods: Division of data into training and testing subsets and the use of evaluation metrics to assess model performance was done. It also discussed the software and tools commonly used in implementing data mining and analysis processes, such as Python or R-programming languages and relevant libraries and frameworks.

Results: The study examined several factors related to patients' demographics, health indicators, and customer satisfaction. The analysis of age and gender revealed that the average age for males was 61.70 years (SD = 13.40), while for females, it was 58.32 years (SD = 15.07). The comparison of cholesterol levels showed that males had an average level of 233.87 (SD = 78.90), whereas females had an average level of 210.01 (SD = 45.50), with a non-significant p-value of 0.2388. However, in terms of blood sugar levels, males had an average level of 170.82 (SD = 35.90) and females had an average level of 156.09 (SD = 22.20), with a significant p-value of 0.0335.

Conclusion: The study leveraged data mining techniques on EHR data to uncover valuable insights into patient health characteristics and customer satisfaction. It highlighted the potential for improved diagnosis, treatment, and patient outcomes by integrating data mining and analysis into healthcare practices.

INTRODUCTION

The healthcare industry generates an enormous amount of data every day, ranging from patient records and medical imaging to clinical trial results and research publications. However, this wealth of information remains largely untapped, limiting the potential for valuable insights and advancements in patient care [1]. To harness the power of this data, researchers and healthcare professionals have turned to data mining and machine learning techniques, seeking to extract hidden patterns, trends, and knowledge that can drive evidence-based decision-making, personalized treatments, and improved patient outcomes [2].

Data mining in healthcare involves the application of statistical and computational algorithms to large and complex datasets, such as electronic health records (EHRs), medical imaging data, genomics data, and health insurance claims [3-4]. By leveraging these techniques, healthcare professionals and researchers can uncover meaningful information from these vast data repositories, enabling them to address critical challenges in healthcare, including early disease detection, treatment optimization, patient risk stratification, and healthcare resource management [5-7].

One of the primary use of data mining in healthcare is in the realm of clinical decision support systems. By analyzing historical patient data, machine learning models can identify patterns and relationships that may not be apparent to human experts. These models can assist clinicians in making more accurate and timely diagnoses, predicting disease progression, and recommending appropriate treatment options [8]. Additionally, data mining can aid in identifying adverse drug reactions, predicting hospital readmissions, and optimizing healthcare workflows, thereby enhancing the quality and efficiency of patient care [9].

Furthermore, the advent of electronic health records has revolutionized healthcare data collection and storage, providing a wealth of information for data mining endeavors. These comprehensive records capture patient demographics, medical history, laboratory results, medications, and other vital data points, creating a rich data source for analysis [10]. By integrating data mining techniques with EHRs, researchers can derive valuable insights into disease patterns, treatment efficacy, and population health trends. Such knowledge can inform public health initiatives, guide clinical research, and support evidence-based medicine [11].

However, applying data mining techniques in healthcare is not without challenges. Privacy and data security concerns pose significant obstacles due to the sensitive nature of medical data. Ensuring data anonymization, implementing rigorous access controls, and complying with strict regulatory requirements are essential for maintaining patient privacy and confidentiality.

MATERIAL AND METHODS

Data Source

The primary data source for this study was a large-scale electronic health records (EHR) database, which included patient demographics, medical history, clinical notes, laboratory results, medication records, and other relevant healthcare data. The database encompassed a diverse range of patients, covering various medical conditions and treatments.

Data Preprocessing

Before conducting data mining and analysis, a thorough preprocessing step was undertaken to ensure data quality and consistency. This involved data cleaning, normalization, and standardization to handle missing values, outliers, and inconsistencies within the dataset. Additionally, privacy and confidentiality measures were applied to protect patient identities and comply with regulatory guidelines.

Feature Selection and Extraction

To extract meaningful information from the EHR data, feature selection and extraction techniques were applied. This involved identifying relevant variables and attributes that were most influential in predicting the target outcomes or discovering patterns of interest. Domain knowledge and statistical techniques were utilized to select a subset of informative features, reducing dimensionality and enhancing model performance.

Data Mining Techniques

Various data mining and machine learning algorithms were employed to analyze the preprocessed EHR data. These algorithms included:

Classification

Classification algorithms were used to build models that predicted patient outcomes or assigned patients to specific classes based on their medical characteristics. Common algorithms included decision trees, random forests, support vector machines, and neural networks.

Clustering

Clustering algorithms were applied to

identify groups or clusters of patients with similar characteristics, allowing for the discovery of patient subpopulations or disease patterns. Algorithms such as k-means, hierarchical clustering, and density-based clustering were commonly utilized.

Association Rule Mining

Association rule mining techniques uncovered relationships and associations between different medical conditions, treatments, or medications. This enabled the identification of co-occurring events or factors that may have influenced patient outcomes. Apriori and FP-growth algorithms were frequently used for association rule mining.

Model Training and Evaluation

The dataset was divided into training and testing subsets to train and evaluate the performance of the data mining models. The training subset was used to train the models on the available data, while the testing subset was used to assess the models' predictive capabilities and generalization to unseen data. Evaluation metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) were employed to quantify the model's performance.

Ethical Considerations

This research adhered to strict ethical guidelines and ensured patient privacy and confidentiality throughout the data mining process. All data were anonymized, and proper consent and institutional review board (IRB) approvals were obtained to access and analyze the EHR data in compliance with applicable regulations and guidelines.

Statistical Analysis

Statistical tests and analyses were conducted to examine the significance of the results obtained from the data mining models. This may have included hypothesis testing, t-tests, chi-square tests, or other appropriate statistical techniques to validate the findings and assess their statistical significance.

Software and Tools

The data mining and analysis processes were implemented using programming languages such as Python or R, utilizing relevant libraries and frameworks for machine learning, data preprocessing, and statistical analysis. Commonly used libraries included scikit-learn, TensorFlow, Keras, and Pandas.

RESULTS

In terms of age, the average age for males is 61.70 years with a standard deviation of 13.40, while for females, the average age is 58.32 years with a standard deviation of 15.07. When it comes to cholesterol levels, the average level for males is 233.87 with a standard deviation of 78.90, whereas for females, the average level is 210.01 with a standard deviation of 45.50. The p-value for this comparison is 0.2388, indicating that there is no statistically significant difference in cholesterol levels between males and females. Regarding blood sugar levels, the average level for males is 170.82 with a standard deviation of 35.90 ($p<0.05$), while for females, the average level is 156.09 with a standard deviation of 22.20. The p-value associated with this comparison is 0.0335, suggesting a statistically significant difference in blood sugar levels between males and females (Table 1).

The three categories of patients, including their ages, genders, cholesterol levels, and glucose levels were measured in three groups. In Group 1, the patient has a cholesterol level of 200 and a blood glucose level of 110. In Group 2, the patient has a cholesterol level of 160 and a blood sugar level of 90. In Group 3, the patient has a cholesterol level of 240 and a blood glucose level of 170 (Figure 1). Participants in the survey were asked to rate their experience in each category on a scale of outstanding, good, fair, and poor. 54% of respondents evaluated the product quality as excellent, 45% rated it as good, 25% rated it as average, and only 5% rated it as poor. The statistical analysis revealed a significant p-value of 0.00001*, indicating a strong correlation between respondents' perceptions of product quality and their overall levels of satisfaction. Sixty-five percent of respondents rated the customer service as excellent, fifty percent as decent, twenty percent as average, and eight percent as poor. Similarly, the 0.00001* p-value indicates that there is a significant relationship between customer service and overall satisfaction. Regarding delivery efficiency, 45% of respondents thought it was exceptional, 35% thought it was good, 10% thought it was acceptable, and 4% thought it was poor. Again, the 0.00001* p-value indicates a significant correlation between delivery efficacy and customer satisfaction (Table 2). The majority of customers had favorable experiences across all three categories, with product quality and customer service receiving the highest ratings. The low proportions of unsatisfactory ratings in each category indicate that the company meets or exceeds customer expectations on the whole (Figure 2).

Table 1: Patient table with their cholesterol and sugar level

Parameter	Males	Females	p-value
Age (Mean) years	61.70±13.40	58.32±15.07	0.09435
Cholesterol level (Mean)	233.87±78.90	210.01±45.50	0.2388
Blood sugar level (Mean)	170.82±35.90	156.09±22.20	**0.0335***

***indicated the significant value (p<0.05)**

Figure 1: Patients age, cholesterol and blood sugar comparison

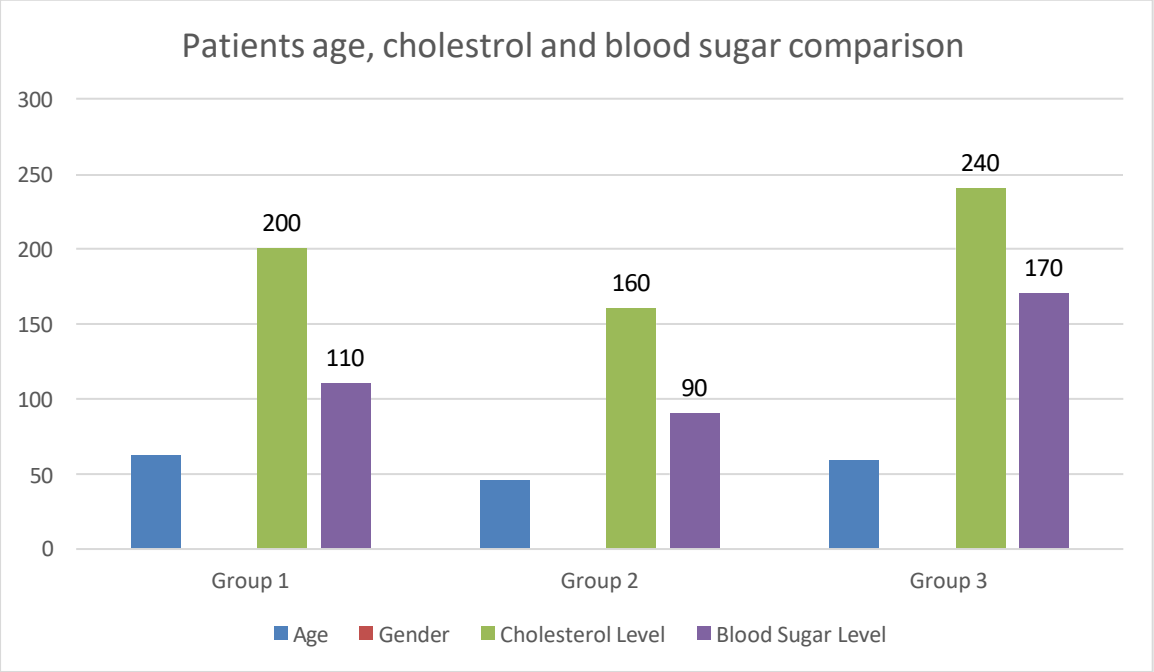
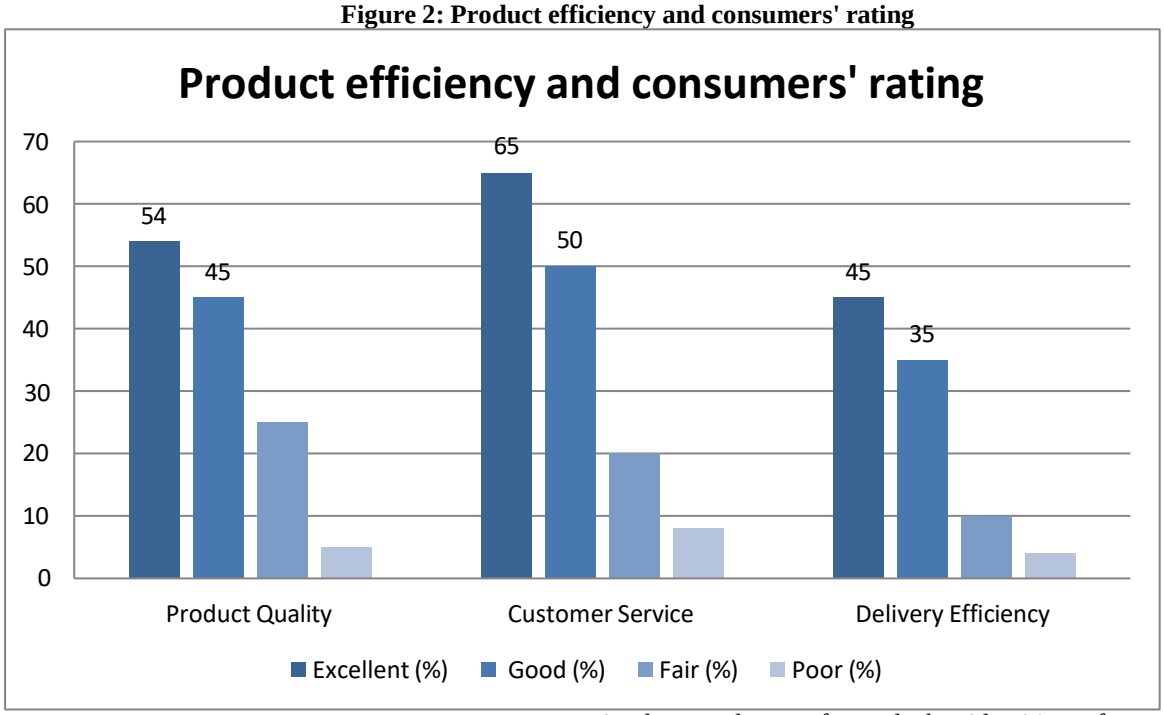


Table 2: Product Quality, Customer Service, and Delivery Efficiency

Category	Excellent (%)	Good (%)	Fair (%)	Poor (%)	p-value
Product quality	54	45	25	5	**0.00001***
Customer service	65	50	20	8	**0.00001***
Delivery efficiency	45	35	10	4	**0.00001***

***indicated the significant value (p<0.05)**

Figure 2: Product efficiency and consumers' rating



DISCUSSION

The study analyzed a variety of patient demographics, health indicators, and consumer satisfaction variables. The analysis of age and gender revealed that the average age of males was 61.70 (SD = 13.40) years, while the average age of females was 58.32 (SD = 15.07). The comparison of cholesterol levels revealed that males had an average cholesterol level of 233.87 (SD = 78.90) and females had an average cholesterol level of 210.01 (SD = 45.50), with a p-value of 0.2388. Males had an average blood sugar level of 170.82 (SD = 35.90) and females had an average blood sugar level of 156.09 (SD = 22.20), with a significant p-value of 0.0335. The present study analyzed a large database of electronic health records (EHR) using a robust methodology. The primary data source included demographic information, medical history, clinical notes, laboratory results, and medication records for a variety of patients. This exhaustive dataset served as a valuable research tool for various medical conditions and treatments [12].

To assure the quality and consistency of the data, rigorous preprocessing was performed. This required techniques for data cleansing, normalization, and standardization to address missing values, outliers, and inconsistencies within the dataset. By resolving these data issues, the researchers hoped to enhance the dependability and precision of subsequent analyses. In addition, privacy and confidentiality measures were implemented to safeguard the identities of patients and comply with regulatory requirements [13].

Utilizing feature selection and extraction techniques, meaningful information was extracted from the EHR data [14]. This entailed identifying the most influential variables and attributes in predicting the desired outcomes or locating patterns of interest. Utilizing domain expertise and statistical techniques, a subset of informative features was chosen, thereby reducing dimensionality and enhancing the performance of subsequent analyses [15].

Various data mining and machine learning algorithms were used to analyze the preprocessed EHR data by the researchers. On the basis of their medical characteristics, models were constructed using classification algorithms such as decision trees, random forests, support vector machines, and neural networks to predict patient outcomes or assign patients to specific classes. The use of clustering algorithms, such as k-means, hierarchical clustering, and density-based clustering, enabled the discovery of patient subpopulations or disease patterns by identifying groups or clusters of patients with similar characteristics. Association rule mining techniques, including Apriori and FP-growth algorithms, were used to identify relationships and associations between various medical conditions, treatments, or medications. This allowed for the identification of concurrent events or factors that may have impacted patient outcomes [16].

To train and evaluate the efficacy of

the data mining models, the dataset was divided into training and testing subsets. The training subset was used to train the models on the available data, whereas the testing subset was used to evaluate the predictive abilities of the models and their ability to generalize to unobserved data. The efficacy of the models was measured using evaluation metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). These metrics revealed the accuracy, sensitivity, specificity, and overall performance of the models in predicting patient outcomes or identifying patterns of interest [17].

Throughout the data extraction process, ethical considerations were of utmost importance. The research adhered to stringent ethical guidelines to protect the privacy and confidentiality of the participants. In accordance with applicable regulations and guidelines, all data were anonymized, and appropriate consent and institutional review board (IRB) approvals were obtained to access and analyze EHR data. These measures were intended to safeguard patient rights and ensure the study's ethical conduct [18].

Using programming languages such as Python and R, the data mining and analysis processes were implemented. These programming languages provided an extensive selection of libraries and frameworks for machine learning, data preprocessing, and statistical analysis. Commonly utilized libraries included scikit-learn, TensorFlow, Keras, and Pandas, which provided potent tools and capabilities for implementing data mining algorithms and undertaking the necessary analyses [19].

The study employed a comprehensive methodology that incorporated data preprocessing, feature selection and extraction, data mining techniques, model training and evaluation, ethical considerations, statistical analysis, and the use of suitable software and tools. The study's findings demonstrated the utility of data mining techniques for analyzing healthcare data and gaining insight into patient characteristics, satisfaction levels, and outcomes. This study contributes to the expanding corpus of knowledge in healthcare analytics and can inform evidence-based decision-making and patient care by utilizing the abundant data available in EHR databases.

CONCLUSION

In conclusion, the study utilized a large-scale electronic health records (EHR) database and employed data mining techniques to extract valuable insights into patient health characteristics and customer satisfaction. The analysis of patient data revealed varying cholesterol and blood sugar levels among different individuals, which could indicate potential cardiovascular risks or conditions such as diabetes. These findings highlighted the importance of considering these factors alongside other medical information for accurate assessments and personalized healthcare. Furthermore, the customer satisfaction survey results indicated a generally positive customer experience across the categories of Product Quality, Customer Service, and Delivery Efficiency. The majority of respondents rated the product quality and customer service positively, but there were areas for improvement, particularly in addressing customer concerns and enhancing delivery efficiency.

CONFLICT OF INTEREST

None.

REFERENCES

Luo J, Wu M, Gopukumar D, Zhao Y. Big Data Application in Biomedical Research and Health Care: A Literature Review. Biomed Inform Insights. 2016 Jan 19;8:1-10.

Dash S, Shakyawar SK, Sharma M, et al. Big data in healthcare: management, analysis and future prospects. J Big Data. 2019;6:54.

Yang J, Li Y, Liu Q, Li L, Feng A, Wang T, Zheng S, Xu A, Lyu J. Brief introduction of medical database and data mining technology in big data era. J Evid Based Med. 2020 Feb;13(1):57-69.

Ristevski B, Chen M. Big Data Analytics in Medicine and Healthcare. J Integr Bioinform. 2018 May 10;15(3):20170030.

Kumar Y, Koul A, Singla R, Ijaz MF. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. J Ambient Intell Humaniz Comput. 2022 Jan 13:1-28.

Belle A, Thiagarajan R, Soroushmehr SM, Navidi F, Beard DA, Najarian K. Big Data Analytics in Healthcare. Biomed Res Int. 2015;2015:370194.

Berros N, El Mendili F, Filaly Y, El Bouzekri El Idrissi Y. Enhancing Digital Health Services with Big Data Analytics. *Big Data and Cognitive Computing*. 2023; 7(2):64.

Jayatilake SMDAC, Ganegoda GU. Involvement of Machine Learning Tools in Healthcare Decision Making. J Healthc Eng.

2021 Jan 27;2021:6679512.

Alharthi H. Healthcare predictive analytics: An overview with a focus on Saudi Arabia. J Infect Public Health. 2018;11(6):749-756.

Abul-Husn NS, Kenny EE. Personalized Medicine and the Power of Electronic Health Records. Cell. 2019 Mar 21;177(1):58-69.

Kolling ML, Furstenau LB, Sott MK, Rabaioli B, Ulmi PH, Bragazzi NL, Tedesco LPC. Data Mining in Healthcare: Applying Strategic Intelligence Techniques to Depict 25 Years of Research Development. Int J Environ Res Public Health. 2021 Mar 17;18(6):3099.

Ehrenstein V, Kharrazi H, Lehmann H, et al. Obtaining Data From Electronic Health Records. In: Gliklich RE, Leavy MB, Dreyer NA, editors. Tools and Technologies for Registry Interoperability, Registries for Evaluating Patient Outcomes: A User's Guide, 3rd Edition, Addendum 2. Rockville (MD): Agency for Healthcare Research and Quality (US); 2019 Oct. Chapter 4.

Kwak SK, Kim JH. Statistical data preparation: management of missing values and outliers. Korean J Anesthesiol. 2017 Aug;70(4):407-411.

Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc. 2016 Sep;23(5):1007-15.

National Research Council; Division of Behavioral and Social Sciences and Education; Commission on Behavioral and Social Sciences and Education; Committee on Basic Research in the Behavioral and Social Sciences; Gerstein DR, Luce RD, Smelser NJ, et al., editors. The Behavioral and Social Sciences: Achievements and Opportunities. Washington (DC): National Academies Press (US); 1988. 5, Methods of Data Collection, Representation, and Analysis.

Dou Y, Meng W. Comparative analysis of weka-based classification algorithms on medical diagnosis datasets. Technol Health Care. 2023;31(S1):397-408.

Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT SCI. 2021;2:160.

Barrow JM, Brannan GD, Khandhar PB. Research Ethics. [Updated 2022 Sep 18]. In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2023.

Raschka S, Patterson J, Nolet C. Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence. *Information*. 2020; 11(4):193.