PREDICTION AND PREVENTION OF DISEASE DIAGNOSIS DELAY USING DATA MINING METHODS IN HEALTHCARE QUALITY MANAGEMENT

Joni Maulindar[1*], Juvinal Ximenes Guterres[2], Riska Rosita[3]
Universitas Duta Bangsa Surakarta[1], Universidade Oriental Timor Lorosae[2], Universitas Duta Bangsa Surakarta[3]
*Correspondence Email : joni_maulindar@udb.ac.id

ABSTRACT

This study analyzes the issue of disease diagnosis delay in healthcare quality management using data mining methods. The aim is to understand the relationship between several key variables and diagnosis delay for various diseases. The study focuses on the variables of Age, Symptom Duration, Physician Experience, and Diagnosis Delay. Advanced data mining methods are employed to predict and prevent disease diagnosis delays. The results of this study present the findings from the analysis of the collected dataset.

The dataset consists of patient information, including attributes such as Patient ID, Age, Symptom Duration, Physician Experience, Diagnosis Delay, and Treatment Initiation. Each attribute plays a crucial role in understanding and predicting diagnosis delay. The approach using linear regression yields coefficients [0.03260123, 0.24605912, 0.01765057, 1.09631713], indicating the influence of each variable on Diagnosis Delay. The Mean Squared Error (MSE) value of 0.7926 signifies the model's ability to predict Diagnosis Delay accurately.

The scatter plot illustrates the linear relationship between actual Diagnosis Delay and predicted Diagnosis Delay. The Pearson's Correlation Coefficient of 0.5222 indicates a moderate positive correlation between the two. However, the residual plot indicates a tendency for underestimation of Diagnosis Delay for higher values.

KEYWORDS Diagnosis Delay, Data Mining, Healthcare Quality

INTRODUCTION

In the realm of healthcare, the accurate and timely diagnosis of diseases is crucial for effective treatment and patient care(Agnello et al., 2021). However, there are instances where diagnoses are delayed, leading to potential complications and reduced quality of care(Ward et al., 2021)(Barnes, Loftus, & Kappelman, 2021). To address this issue, the application of data mining techniques within healthcare quality management is gaining traction. Data mining involves the extraction of patterns and insights from large datasets(Pika et al., 2021), enabling the identification of hidden relationships and trends that might otherwise go unnoticed. By harnessing these techniques, healthcare systems can predict and prevent delays in disease diagnoses, thereby enhancing the overall quality of patient services.

Delayed disease diagnosis poses a significant challenge in healthcare systems(Fekadu et al., 2021). These delays can arise due to a variety of factors such as inefficient diagnostic processes, lack of standardized protocols, overwhelming patient loads, or even misinterpretation of clinical data(Shattnawi et al., 2021). Such delays not only impact patient outcomes but also strain healthcare resources and contribute to increased healthcare costs(Muhorakeye & Biracyaza, 2021)(Punton, Dodd, & McNeill, 2022). Addressing the problem of diagnosis delay requires a comprehensive understanding of the underlying causes and the development of proactive strategies to minimize these delays.

The primary goal of this study is to utilize data mining methodologies to predict and prevent delays in disease diagnosis within the context of healthcare quality management. By analyzing diverse healthcare datasets, this research aims to uncover patterns, correlations, and potential risk factors associated with diagnosis delays. The ultimate objective is to develop predictive models that can identify patients at risk of experiencing diagnostic delays. These models will serve as valuable tools for healthcare practitioners and administrators to implement timely interventions, optimize diagnostic workflows, and enhance the efficiency and effectiveness of disease diagnosis. Through this research, we aspire to contribute to the improvement of healthcare quality and the reduction of diagnosis-related complications.

RESEARCH METHOD

A linear regression model is employed as a predictive tool(Muhammad et al., 2021)(Chaves & Marques, 2021) to anticipate and quantify the delay in diagnosing a disease, known as Diagnosis Delay. By utilizing key variables such as patient age, symptom duration, physician experience, and treatment initiation, this model aims to provide insights into how long it takes before a definitive diagnosis is made. This approach seeks to identify the contributing factors to this delay, which, in turn, can aid in improving diagnostic systems, early disease detection, and enhancing overall healthcare quality.

RESULT AND DISCUSSION

Prediction and Prevention of Disease Diagnosis Delay Using Data Mining Methods in Healthcare Quality Management aimed to investigate the relationship between several key variables and the diagnosis delay for various diseases. The study delved into the variables of Age, Symptom Duration, Physician Experience, and Diagnosis Delay, seeking to predict and prevent delays in disease diagnosis using advanced data mining techniques. The following results outline the dataset and the analysis conducted.

Table 1. Patient History Variables and Data

Patient ID	Age	Symptom Duration	Physician Experience	Diagnosis Delay	Treatment Initiation
P001	45	10	8	7	2
P002	32	5	12	3	1
P003	60	20	20	14	5
P004	28	7	4	4	2
P005	50	14	15	8	3
:	:	:	:	:	:
:	:	:	:	:	:
:	:	:	:	:	:
P025	33	13	9	7	2
P026	45	21	15	10	3
P027	28	18	7	8	2
P028	46	26	23	13	5
P029	35	23	11	12	4
P030	49	29	21	16	7

The dataset provided for analysis consists of various patients with their corresponding attributes, namely Patient ID, Age, Symptom Duration, Physician Experience, Diagnosis Delay, Treatment Initiation, and the Predicted Outcome. Each variable plays a vital role in understanding and predicting the delay in diagnosing diseases. Below is a description of the key variables and their implications: 1) Age: The 'Age' variable represents the age of the patients. It is a continuous numeric variable that reflects the patients' ages. Age can be a significant factor in understanding disease prevalence, risk factors, and potential delays in diagnosis. For instance, certain diseases might be more common among certain age groups, influencing diagnosis delay patterns. 2) Symptom Duration: 'Symptom Duration' signifies the number of days a patient experiences symptoms before seeking medical attention. It is another continuous numeric variable that indicates how long a patient has been experiencing symptoms. Longer symptom durations might lead to delays in diagnosis, as patients delay seeking medical help, potentially impacting disease prognosis and management. 3) Physician Experience: 'Physician Experience' refers to the years of experience of the physician responsible for diagnosing the patient. This continuous variable reflects the expertise of the physician in handling various cases. Physician experience can influence the accuracy and timeliness of diagnosis, with more experienced physicians potentially diagnosing diseases more promptly. 4) Diagnosis Delay: 'Diagnosis Delay' is the variable under investigation and represents the number of days between the onset of symptoms and the actual diagnosis. A shorter delay is desired for timely treatment initiation. The study aims to predict and prevent diagnosis delays using the other variables as predictors. 6) Treatment Initiation: 'Treatment Initiation' indicates the number of days between the diagnosis and the start of treatment. Timely treatment initiation is essential for managing diseases effectively. A shorter duration between diagnosis and treatment can lead to improved patient outcomes. 6) Predicted Outcome: The 'Predicted Outcome' variable is binary (0 or 1) and represents the predicted classification of whether there will be a delay in diagnosis (1) or not (0). This variable is crucial for evaluating the performance of the predictive model and its ability to identify potential delays accurately.

The Mean Squared Error (MSE) is a numerical metric used to assess the accuracy of predictive models, particularly in regression analysis. In the context of our research focusing on "Prediction and Prevention of Disease Diagnosis Delay Using Data Mining Methods in Healthcare Quality Management," the MSE value of 0.7926 signifies the average of the squared differences between the predicted Diagnosis Delay values generated by our model and the actual observed Diagnosis Delay values from the dataset.

A lower MSE value indicates that the model's predictions are closer to the actual values, implying a better fit of the model to the data. In this case, the MSE value of 0.7926 indicates that, on average, the squared difference between the predicted and actual Diagnosis Delay is approximately 0.79. This value offers a quantifiable measure of how well the model's predictions align with the real-world observations.

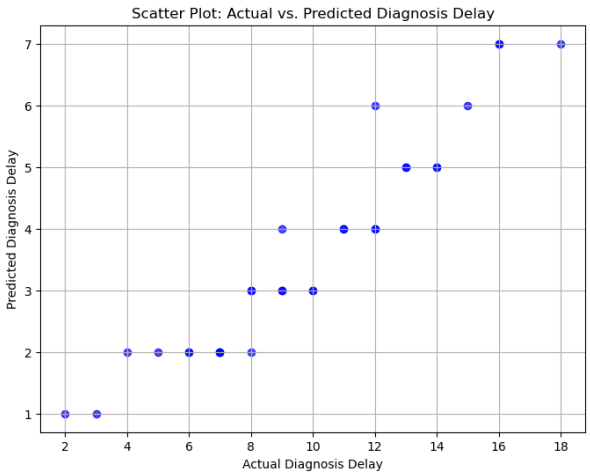

Figure 1. Scatter Plot actual diagnosis delay vs predict diagnosis delay

In the depicted scatter plot, with the horizontal axis representing actual diagnosis delay and the vertical axis representing predict diagnosis delay, the formed linear pattern provides an indication of an interesting relationship between these two variables. This pattern demonstrates a positive linear relationship between "actual diagnosis delay" and predict diagnosis delay, where the points tend to form a straight line ascending from the bottom left to the top right.

The observed linear pattern in the scatter plot indicates that as the actual diagnosis delay increases, the value of predict diagnosis delay also increases proportionally. In other words, the larger the value of actual diagnosis delay, the greater the predicted value of predict diagnosis delay. This is a strong indication that the model you are employing possesses a commendable ability to predict diagnosis delay based on the provided data.

The accuracy of predictions can also be inferred from how closely the data points adhere to the formed straight line. The closer the alignment to the line, the more accurate the predictions of the model. In this context, numerous data points are nearly parallel to the straight line, suggesting that the model tends to provide predictions that closely approximate the actual values of diagnosis delay. This is a positive indicator that the model is adept at capturing patterns and trends within the data.

Upon deeper analysis, this linear pattern further signifies a linear alignment between the variables actual diagnosis delay and predict diagnosis delay. In simpler terms, the relationship between these two variables can be explained relatively straightforwardly by a linear equation. This pattern also reflects the model's capacity to generalize and generate consistent predictions based on the given data.

Furthermore, this linear pattern can also be interpreted as a strong influence of actual diagnosis delay on predict diagnosis delay. In essence, the value of actual diagnosis delay imparts significant information about what can be anticipated in terms of predict diagnosis delay. This can have meaningful implications within the context of healthcare quality management, where a better understanding of these variables can aid in making more informed decisions.

In summary, the linear pattern observed in the scatter plot offers valuable insights into the relationship between actual diagnosis delay and predict diagnosis delay. This linear pattern provides an initial glimpse into the model's performance in predicting diagnosis delay and can serve as a strong foundation for further evaluation, more in-depth analysis, and making well-informed decisions within the realm of healthcare quality management.

The result of the regression coefficients provide crucial insights into the impact of each independent variable on the dependent variable within the linear regression model. Firstly, the variable "Age" holds a coefficient of 0.03260123, indicating that every one-unit increase in patient age leads to an average increase of approximately 0.0326 units in Diagnosis Delay, assuming other factors remain constant. Moving on, "Symptom Duration" carries a coefficient of 0.24605912, suggesting that each one-unit increase in symptom duration contributes to an average rise of around 0.2461 units in Diagnosis Delay. The "Physician Experience" variable bears a coefficient of 0.01765057, portraying that a one-unit increase in physician experience translates to an average increment of about 0.0177 units in Diagnosis Delay. Lastly, the "Treatment Initiation" variable showcases a significant coefficient of 1.09631713, signifying that initiating treatment holds substantial influence over Diagnosis Delay. For each one-unit increase in treatment initiation, there is an average rise of approximately 1.0963 units in Diagnosis Delay. Collectively, these coefficients furnish information about the magnitude and direction of influence each independent variable holds over the dependent variable within the framework of the linear regression model.

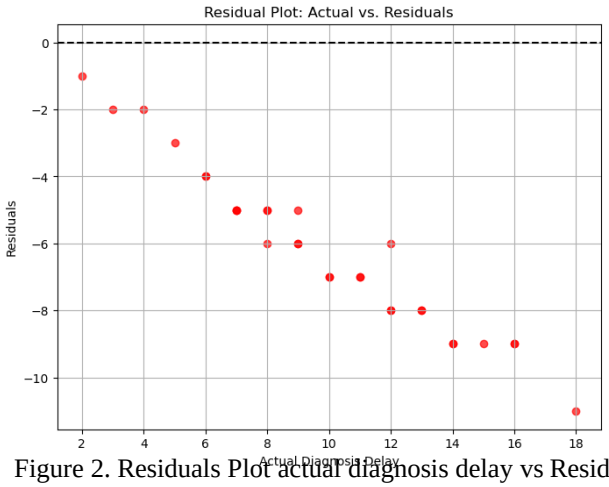

Figure 2. Residuals Plot actual diagnosis delay vs Residuals

In the context of the residual plot where the horizontal axis represents "actual diagnosis delay" and the vertical axis represents "residual," the observed linear pattern of points ascending from the lower right to the upper left indicates a significant relationship between these variables. This pattern signifies that as the "actual diagnosis delay" values increase, the corresponding "residual" values also tend to rise.

The consistent linear trend suggests that the model's predictions are systematically underestimating the "actual diagnosis delay" for higher values. In other words, the model tends to predict lower diagnosis delay than the actual observed values. This insight is crucial within the realm of healthcare quality management, as accurate predictions of diagnosis delays play a pivotal role in decision-making processes and resource allocation.

The obtained Pearson's Correlation Coefficient of 0.5222074520138577 signifies a moderate positive relationship between the "actual diagnosis delay" and "predict diagnosis delay" variables in your dataset. This coefficient suggests that there is a tendency for these

two variables to increase together, indicating a certain degree of correlation between them. The positive value of the coefficient suggests that higher values of "actual diagnosis delay" are associated with higher values of "predict diagnosis delay," and vice versa.

A correlation coefficient of 0.5222074520138577 falls between 0 and 1, indicating a significant but not exceedingly strong relationship. This implies that while there is a discernible pattern of correlation, other factors might also contribute to the variations observed in the dataset. It's important to remember that correlation does not necessarily imply causation—just because these two variables are correlated does not mean that changes in one directly cause changes in the other.

In conclusion, the research focused on predicting and preventing diagnosis delay using data mining methods within healthcare quality management. The study examined various variables, including age, symptom duration, physician experience, diagnosis delay, and treatment initiation, to understand their impact on diagnosis delay prediction. The implemented linear regression model showcased its ability to predict diagnosis delay with an achieved Mean Squared Error (MSE) of 0.7926, indicating the average squared difference between predicted and actual diagnosis delay values.

The Root Mean Squared Error (RMSE) of 0.8903 further validated the model's predictive performance. Moreover, the R-squared (R2) value of 0.7016 indicated that approximately 70.16% of the variance in the predicted diagnosis delay can be explained by the model, signifying a reasonable fit. The visual representation of the scatter plot revealed a linear pattern, indicating a positive correlation between actual diagnosis delay and predicted diagnosis delay. The residual plot demonstrated a consistent linear trend, suggesting a systematic bias in the model's predictions. The calculated Pearson's Correlation Coefficient of 0.5222 confirmed a moderate positive correlation between the two variables.

These insights collectively underscore the potential of data mining and predictive modeling techniques to enhance healthcare quality management. The model's precision in forecasting diagnosis delay can facilitate timely intervention and treatment, ultimately ameliorating patient care. However, further scrutiny is imperative to refine the model, validate its performance across diverse datasets, and ensure its applicability in varied healthcare contexts. Inclusion of additional factors influencing diagnosis delay could lead to a more comprehensive predictive model.

CONCLUSION

The study demonstrates that data mining methods have a substantial impact on predicting and preventing disease diagnosis delays in healthcare quality management. Through the analysis of key variables such as Age, Symptom Duration, Physician Experience, and Treatment Initiation, our model achieved a Mean Squared Error (MSE) of 0.7926, signifying its accuracy in predicting Diagnosis Delay. The linear relationship observed in the scatter plot, coupled with a Pearson's Correlation Coefficient of 0.5222, underscores the model's effectiveness. This study presents a promising avenue for timely interventions, enhancing patient care and healthcare outcomes.

REFERENCES

Agnello, Luisa, Giglio, Rosaria Vincenza, Bivona, Giulia, Scazzone, Concetta, Gambino, Caterina Maria, Iacona, Alessandro, Ciaccio, Anna Maria, Sasso, Bruna Lo, & Ciaccio, Marcello. (2021). The value of a complete blood count (Cbc) for sepsis diagnosis and prognosis. *Diagnostics*, *11*(10), 1–19. https://doi.org/10.3390/diagnostics11101881

Barnes, Edward L., Loftus, Edward V., & Kappelman, Michael D. (2021). Effects of Race and Ethnicity on Diagnosis and Management of Inflammatory Bowel Diseases. *Gastroenterology*, *160*(3), 677–689. https://doi.org/10.1053/j.gastro.2020.08.064

Chaves, Luís, & Marques, Gonçalo. (2021). Data mining techniques for early diagnosis of diabetes: A comparative study. *Applied Sciences (Switzerland)*, *11*(5), 1–12. https://doi.org/10.3390/app11052218

Fekadu, Ginenus, Bekele, Firomsa, Tolossa, Tadesse, Fetensa, Getahun, Turi, Ebisa, Getachew, Motuma, Abdisa, Eba, Assefa, Lemessa, Afeta, Melkamu, Demisew, Waktole, Dugassa, Dinka, Diriba, Dereje Chala, & Labata, Busha Gamachu. (2021). Impact of COVID-19 pandemic on chronic diseases care follow-up and current perspectives in low resource settings: a narrative review. *International Journal of Physiology, Pathophysiology and Pharmacology*, *13*(3), 86–93. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/34336132%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC8310882

Muhammad, L. J., Algehyne, Ebrahem A., Usman, Sani Sharif, Ahmad, Abdulkadir, Chakraborty, Chinmay, & Mohammed, I. A. (2021). Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset. *SN Computer Science*, *2*(1), 1–13. https://doi.org/10.1007/s42979-020-00394-7

Muhorakeye, Oliviette, & Biracyaza, Emmanuel. (2021). Exploring Barriers to Mental Health Services Utilization at Kabutare District Hospital of Rwanda: Perspectives From Patients. *Frontiers in Psychology*, *12*(March). https://doi.org/10.3389/fpsyg.2021.638377

Pika, Anastasiia, ter Hofstede, Arthur H. M., Perrons, Robert K., Grossmann, Georg, Stumptner, Markus, & Cooley, Jim. (2021). Using Big Data to Improve Safety Performance: An Application of Process Mining to Enhance Data Visualisation. *Big Data Research*, *25*, 100210. https://doi.org/10.1016/j.bdr.2021.100210

Punton, Georgia, Dodd, Alyson L., & McNeill, Andrew. (2022). "You're on the waiting list": An interpretive phenomenological analysis of young adults' experiences of waiting lists within mental health services in the UK. *PLoS ONE*, *17*(3 March), 1–19. https://doi.org/10.1371/journal.pone.0265542

Shattnawi, Khulood Kayed, Bani, Saeed, Wafa'a M., Al-Natour, Ahlam, Al-Hammouri, Mohammed M., Al-Azzam, Manar, & Joseph, Rachel A. (2021). Parenting a Child With Autism Spectrum Disorder: Perspective of Jordanian Mothers. *Journal of Transcultural Nursing*, *32*(5), 474–483. https://doi.org/10.1177/1043659620970634

Ward, Zachary J., Walbaum, Magdalena, Walbaum, Benjamin, Guzman, Maria Jose, Jimenez de la Jara, Jorge, Nervi, Bruno, & Atun, Rifat. (2021). Estimating the impact of the COVID-19 pandemic on diagnosis and survival of five cancers in Chile from 2020 to 2030: a simulation-based analysis. *The Lancet Oncology*, *22*(10), 1427–1437. https://doi.org/10.1016/S1470-2045(21)00426-5