Modeling of Machine Learning for Analyzing Key Patterns in Healthcare Data: A Data Mining Analysis

***Abstract*— This study investigates the use of data mining and machine learning techniques to analyze healthcare data and identify factors influencing medical costs and hospital admissions. Leveraging a dataset of 55,500 patient records, the research applies preprocessing, feature engineering, and classification modeling to reveal critical patterns in demographics, medical conditions, and billing behavior. A Decision Tree model was used to classify billing categories and analyze admission types, with a focus on optimizing hospital operations and improving resource management. The model's performance was improved through stratified sampling, ensuring balanced representation across diagnostic outcomes. Insights from this analysis provide actionable recommendations for enhancing financial planning, hospital efficiency, and patient care strategies through data-driven decision-making.**

***Keywords*— decision tree; Healthcare analytics; Medical costs; Data mining; Hospital admissions; Predictive modeling**

I. INTRODUCTION

In today's data-driven world, healthcare organizations increasingly rely on data analysis techniques to enhance decision-making, optimize resource allocation, and improve patient care. The rapid growth of healthcare data presents both opportunities and challenges, as extracting meaningful insights from large datasets requires structured analysis and effective methodologies. Understanding the key factors influencing medical costs and patient admissions is crucial for improving hospital management, optimizing financial planning, and ensuring efficient use of resources.

This project focuses on exploring healthcare data to identify patterns and relationships affecting medical treatment costs and hospital admissions. The dataset includes attributes such as patient demographics (age, gender, blood type), medical conditions, hospital details, admission types, insurance providers, and billing amounts. By applying various data analysis techniques, we seek to uncover hidden trends that can help explain variations in medical costs, classify admission types, and analyze the impact of different factors on hospital expenses.

One of the main challenges in healthcare analytics is ensuring data quality and handling complex relationships between multiple variables. To address this, we will preprocess the data by handling missing values, standardizing numerical attributes, and selecting the most relevant features. We will also explore different analytical approaches to determine the most effective methods for generating meaningful insights.

By the end of this project, our findings will contribute to the growing field of healthcare analytics, demonstrating how data-driven approaches can optimize hospital operations, improve financial planning, and enhance patient care services.

II. LITERATURE REVIEW

A. Rapid Modelling of Machine Learning

Rapid modelling has been a long-time concern by

Promising, and it is also one of the suggested algorithms from AutoModel used in this research together with Decision Tree [20] and Support Vector Machine [21]. To the best of our knowledge, rapid modelling on office rental prediction has not been reported yet in the current literature. This research filled the gap by presenting the precise steps and the comparison of results.

B. End-of-Life Medical Costs in Cancer Care

End-of-life (EOL) care for older adults with cancer has been widely studied due to its high cost and limited clinical benefit. Jo et al. (2023) analyzed 6,098 cases in South Korea using national insurance data and found that over 60% of medical expenses occurred in the last three months of life. The study highlights the financial burden of high-intensity treatments and suggests that early palliative care and policy measures may reduce unnecessary costs and improve patient quality of life.

c. Oral Health Knowledge and perception Among Diabetes Mellitus Patients

Diabetes mellitus is a growing global health issue that also affects oral health. This study assessed oral health knowledge among 120 diabetic patients using structured interviews. Results showed that most participants had limited knowledge and poor dental care habits, despite regular tooth brushing. While 74.2% recognized the link between diabetes and tooth decay, and 89.2% acknowledged its connection to gum disease, many still lacked awareness. The study highlights the need for better education, improved oral hygiene practices, and easier access to dental care for diabetic patients.

III. DATASET DESCRIPTION:

OVERVIEW:

This healthcare dataset contains **55,500 patient records**, each comprising demographic, clinical, and administrative information. Core features include **age**, **gender**, **blood type**, and **medical condition**, as well as hospital-related data such as **date of admission**, **discharge date**, **doctor**, **hospital**, and **insurance provider**. The **gender distribution is nearly perfectly balanced,** with **27,774 males** and **27,726 females**, making the dataset unbiased in terms of sex-based analysis. The **blood types** are also evenly distributed, with eight major types represented, including **A-, A+, B-, B+, AB-, AB+, O-, and O+**, each with nearly equal counts (~6,900 each), ensuring good variability. Financial and logistical attributes such as **billing amount** and **room number** are included, and **medication** data adds pharmacological context. The target variable, **Test Results**, is split into three diagnostic categories: *Abnormal*, *Normal*, and *Inconclusive*. This dataset is well-suited for training classification models due to its balanced class distributions and diverse set of structured features, both numerical and categorical.

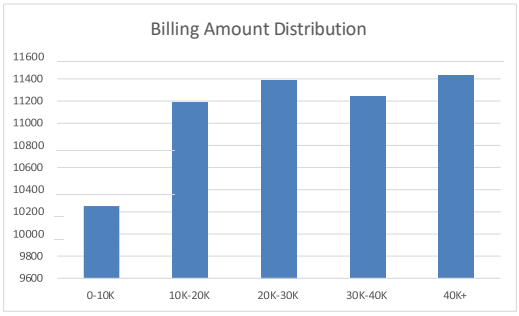

Figure 1: Visualizes how patients are distributed across billing cost ranges.

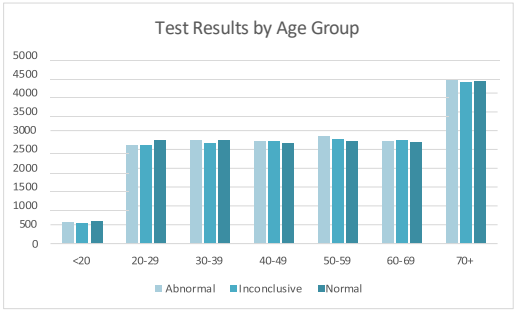

Figure2: shows how test results vary by age group, highlighting which ages are more likely to have abnormal or inconclusive outcomes

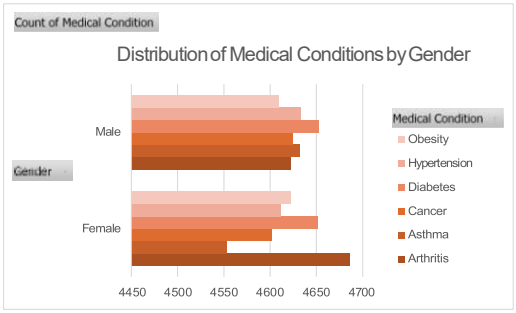

Figure 3: illustrates the gender-based distribution of chronic medical conditions, providing insights into potential sex-specific prevalence patterns across the dataset

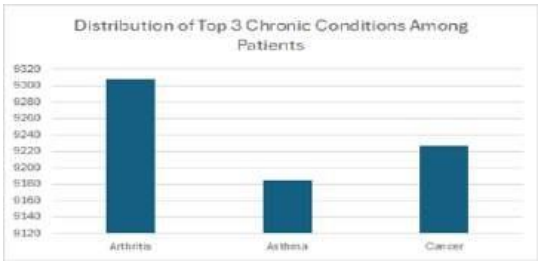

Figure 4: The three most common chronic conditions among all patients

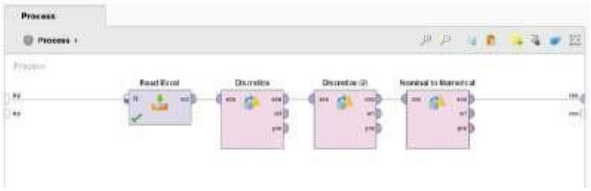

Figure5: Initial preprocessing setup including data reading and basic transformation tools in RapidMiner.

B. Feature Engineering

To enrich the dataset with relevant insights, a new attribute, `Stay Duration`, was generated using the `date_diff ()` function, representing the number of days between "Date of Admission" and "Discharge Date". This transformation allowed us to quantify hospitalization periods, a potentially significant factor in the analysis phase.

Figure6: Use of the "Generate Attributes" operator to create a new feature (Stay Duration) based on admission and discharge dates.

IV. Data Preprocessing and model building:

Data preprocessing is a critical step in the knowledge discovery process, ensuring that the dataset is suitable for modeling and analysis. The dataset used in this project was obtained in Excel format and consisted of approximately 55,501 records and 15 attributes related to patient information, medical conditions, and billing details. Key attributes included "Patient Name", "Age", "Gender", "Medical Condition", "Admission Type", "Billing Amount", and "Discharge Date".

A. Data Cleaning:

preliminary review of the dataset confirmed that there were no missing values, duplicate records, or obvious outliers. This facilitated a more direct transition to the data transformation stage.

C. Data Normalization:

Numerical attributes, particularly "Age" and "Billing Amount", exhibited varying scales. To prepare the data for algorithms sensitive to magnitude differences, we applied normalization techniques using the "Normalize" operator in RapidMiner. This step helped standardize value ranges and improve the reliability of distance-based or gradient-based models.

V. Detailed Workflow Description

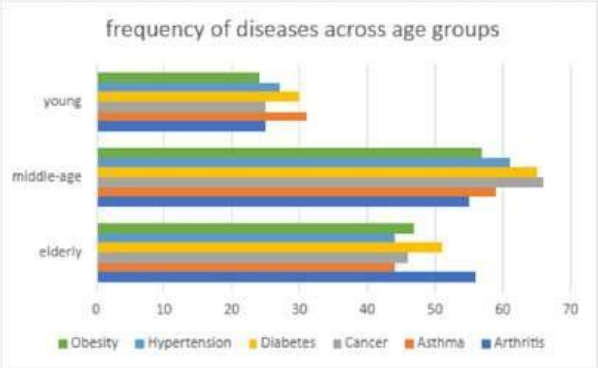

Figure 7: We conclude the elderly are the most susceptible to diseases such as hypertension and diabetes, while obesity cases are higher in the middle-aged group.

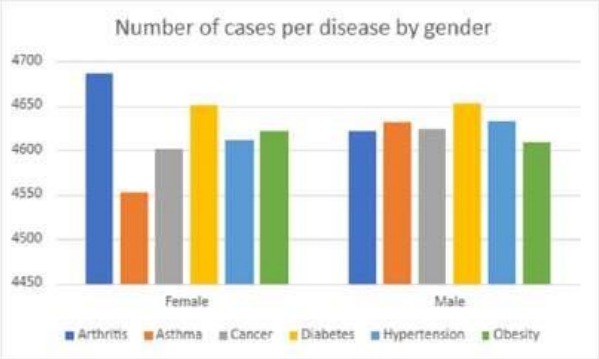

Figure 8: We conclude Women are more affected by arthritis and asthma, while Men are more prone to diabetes and hypertension.

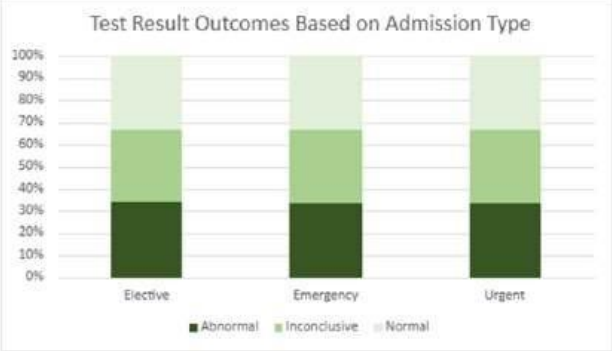

Figure 9: Test results vary by admission type, with abnormal results being higher in emergency admissions, while normal results are more common in elective admissions.

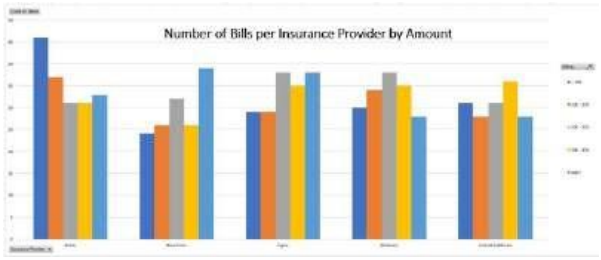

Figure 10: There is a clear variation in billing amounts among insurance providers, indicating that different insurers are associated with different financial burdens, possibly due to variations in coverage or patient types.

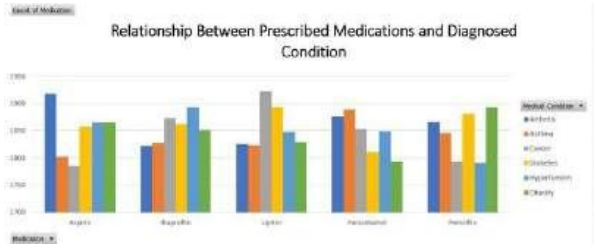

Figure 11: There is a noticeable pattern linking specific medications to medical conditions, suggesting the presence of standardized prescribing practices and condition-specific treatment protocols.

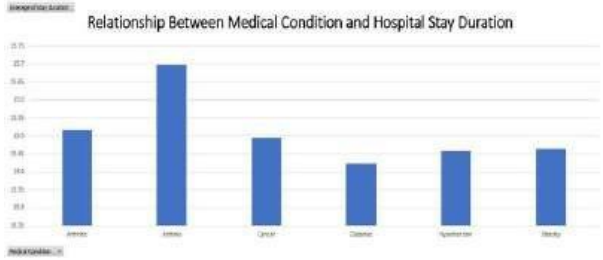

Figure 12: Some medical conditions lead to longer hospital stays than others. This shows that the type of illness affects how long patients need to stay in the hospital.

VI. SELECTED DATA MINING TECHNIQUE

For our analysis, we will choose one primary data mining technique to complement the project's objectives:

• Classification: will be used to group records in the dataset into predefined categories based on their attributes. In this project, the main goal will be to predict billing amounts (e.g., <10K, 10K–20K, 30K–40K, or >40K) based on patient and hospital-related information such as age, gender, insurance provider, and stay duration. By applying classification algorithms, we will aim to identify meaningful patterns in the data that help us understand how different patient characteristics influence healthcare costs. To achieve this, we will use a Decision Tree model in RapidMiner. The model will help visualize the most important decision-making factors. Attributes like age, insurance provider, and stay duration will be shown to play key roles in determining billing amounts. This method will allow healthcare providers to anticipate billing outcomes and take proactive steps to manage healthcare costs more efficiently. It will also support more informed decision-making in hospital operations by identifying which factors most influence patient

billing. When combined with attribute selection and proper visualization, classification will become a powerful tool for revealing hidden trends and improving financial planning in healthcare service delivery.

VII. MODEL EVALUATION

Due to significant class imbalance in the dataset, early model iterations using the full sample failed to detect minority classes, resulting in extremely skewed predictions toward the Abnormal class only.

To resolve this, **stratified sampling** was applied to create a more balanced training set.

After testing multiple sample sizes, a size of 700 was found to yield the best balance.

With this configuration, the Decision Tree model was able to successfully classify all three target labels. The performance notably improved, achieving an accuracy of 32.67%, a weighted recall of 32.93%, and weighted precision of 15.56%.

Most importantly, the model showed meaningful predictions across all classes, unlike earlier results where Normal and Inconclusive classes were completely ignored.

Figure 13: **Accuracy**
(*32.67% ± 3.44)*

Figure 14: **Precision**
(15.56% ± 10.93)

Figure 15: **Recall**
(32.93% ± 3.55)

REFERENCES

[1] A. Abuaisha, F. A. Eshumani, K. Y. Elshoumani, and H. S. Eshtiwi, “Oral Health Knowledge and Perception Among Diabetes Mellitus Patients,” NAJSP, vol. 2, no. 4, pp. 140–144, 2024.

[2] J. Brownlee, "Machine Learning Mastery With Weka," 1st ed., 2019.

[3] M. W. Hasan, M. A. Razaque, and A. S. M. Kayes, "Healthcare Data Mining Techniques: A Review and Case Study," *IEEE Access*, vol. 11, pp. 44683–44697, 2023. doi: 10.1109/ACCESS.2023.3247075

[4] L. Hu et al., "Healthcare Cost Prediction and Analysis Using Data Mining Techniques," *Springer*, vol. 10, no. 2, pp. 215–230, 202