KINGDOM OF SAUDI ARABIA
Ministry of Education
Taibah University
College of Computer Science and Engineering
(Female Section)

جامعة طيبة
TAIBAH UNIVERSITY

DS331 – Data Mining & Analytics 1

Project by:

Eithar Shehabuddin Yanj | 4450113
Layan Hulayyil Aljuhani | 4456556
Manar Abdulmohsen Alaoufi | 4456513
Manar Murdi Aldosari | 4456086

Section: F1

Supervised by

Dr. Samar Alsaleh
Semester 2 (2024/2025)

Project description :

The dataset "AI in Healthcare" is a synthetic collection of 5000 rows and 20 columns, meticulously designed for academic research and analysis in the field of healthcare and artificial intelligence (AI). Each column in the dataset is purposefully named to reflect its relevance to healthcare-related aspects, enabling comprehensive investigations into the intersection of AI and healthcare.

The dataset encompasses diverse healthcare attributes, such as patient demographics (age, gender), vital signs (blood pressure, heart rate, temperature), medical diagnosis, prescribed medications, treatment durations, insurance types, attending physician information, hospital affiliations, lab test results, X-ray outcomes, surgical procedures, recovery times, patient allergies, family medical histories, patient satisfaction scores, and AI-assisted diagnosis confidence levels.

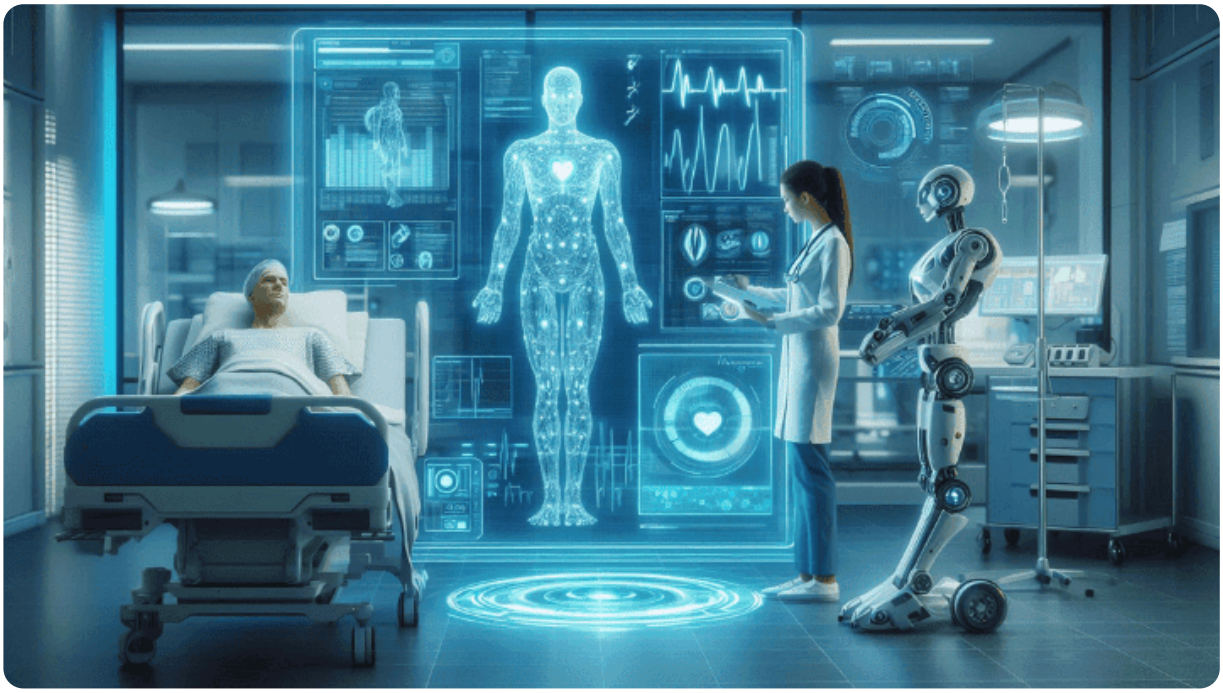
This case study explores clinical patient data to uncover insights into health indicators, diagnoses, medication patterns, and healthcare access. Using descriptive statistics (measures of central tendency, dispersion, and association) and visualization tools, we aim to:
- Analyze patterns in patient vitals, diagnoses, and treatments.
- Identify relationships between symptoms, risk factors, and outcomes.

Contribution report :

Our team collaborated and worked together efficiently through hard teamwork. Each member was assigned one main role for each lab, then together learned each others task and summarized the achievements.
Roles were assigned as the following:

Eithar Yanj : Uniqueness,Validity, Relevance,Handling Outliers, Generating Association Rules (Step 8+9), Decision Tree Classifier.
Layan Aljuhani : Measures of Central Tendency, Accuracy Dimension, Handling Data-Type Issues, Mining Frequent Patterns, Preparing Data for Classification.
Manar Alaoufi : Measures of Association, Consistency, Handling Inconsistent, Generating Association Rules (Step 7), Visualizing Association Rules as a Network Graph3, Naive Bayes Classifier.
Manar Aldosari : Measures of Dispersion, Completeness, Handling Missing Values, Preparing Data, Documenting Findings from Pattern Mining, Comparing Results Across Models.

Lab #1 – Selecting and Validating a Data Mining Project



Step 1: Identify a Real-World Problem

Problem:
Our project aims to analyze the accuracy of AI diagnostics to improve the quality of healthcare services provided to patients.

What challenge or business question do I want to address?

1. Ratio of confidence in diagnosis?
2. How can data be leveraged to increase diagnostic accuracy?
3. What is the difference between human and AI diagnosis?
4. What is the impact or potential risks depending on the accuracy of AI diagnosis?

Is there value in solving this problem?
Yes, quickly detect the disease and receive health care, accurately increase the patient's diagnosis and avoid medical errors as much as possible

Who are the stakeholders who will benefit from the insights?
Many will benifit such as:

1. Ministry of Health
2. Healthcare Facilities (Doctors, Nurses)
3. Patients

Step 2: Ensure the Problem is Data-Driven Your project must rely on data

Is there relevant publicly available data?
Yes, the dataset includes information about patients, their diagnosis, treatment, and the results of medical tests they underwent.

Is the data structured and suitable for analysis?
Yes, the data is structured and analyzable, containing 5,000 rows and 20 columns, making it easy for us to use techniques to analyze its data

Is the dataset large enough for meaningful analysis?
Yes, it is enough for sufficient to extract relevant results that have statistical value.

Step 3: Check if it Fits the CRISP-DM Framework

Alignment with the six phases of CRISP-DM:

1. Business Understanding – Can you clearly define the problem and objective?
The problem is clearly defined: measuring the accuracy of AI diagnoses and assessing patient satisfaction with the services provided.

2. Data Understanding – Is the data available and relevant?
The data can be effectively cleaned and processed, including handling missing or outlier values.

3. Data Preparation – Can the data be cleaned and prepared effectively?
The data can be effectively cleaned and processed, including handling missing or outlier values.

4. Modeling – Can you apply frequent pattern mining, classification, and clustering?
Key data mining techniques will be applied, such as:
• Classification: to predict diagnoses based on patient characteristics.
• Clustering: to discover patterns among patients based on medical factors.
• Pattern Extraction: to identify common trends among different disease conditions.

5. Evaluation – How do you measure success?
The success of the project will be measured by:
• Comparing the diagnostic accuracy of AI and human diagnoses.
• Analyzing patient satisfaction with the treatments they received.

6. Deployment – Can the insights be used in real-world applications?
The results can be used to improve AI models in healthcare and enhance confidence in AI-based diagnostic systems.

Step 4: Check Project Feasibility

Does the project address a meaningful real-world problem?

✅Yes, as it will help may people!

Will the results provide actionable insights?

✅Yes

Can you find a publicly available dataset that fits your idea?

✅Yes, we found ours on Kaggle.com

Is the dataset large enough to support analysis but small enough to handle within the course timeline?

✅Yes, perfect size!

Are the features in the dataset relevant to your problem?

✅Yes as we can use it for all tasks and phases needed.

Is the project scope manageable within the time and resources available?

✅Yes

Can you apply all three key data mining techniques to the dataset?

✅Yes