PROBLEM STATEMENT:

Knowledge Representation and Insight Generation from Structured Datasets

AN INTEL UNNATI PROJECT

ORGANISATION : INTEL UNNATI INDUSTRIAL

TEAM NAME : DEV MASTERS

MEMBERS : TEJASWEE KUMAR SINGH
AMIT KUMAR BEHERA

MENTOR : SUBRATA KUMAR MOHANTY

INSTITUTE : INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY, BHUBANESWAR

Introduction

INTRODUCTION

The explosion of data in today's world presents a double-edged sword. While we have access to more information than ever before, this data often resides in structured formats like tables, spreadsheets, or databases. These formats, while organized, can be rigid and lack the context needed for traditional AI techniques to truly understand the information.

OBJECTIVES

- Develop methods to effectively represent knowledge from structured datasets.
- Design algorithms to analyze and reason over this knowledge representation
- Generate insights and uncover hidden patterns within the data.

Real World Impact

- Empowering informed decision-making: Extracted knowledge can be used for tasks like risk assessment, fraud detection, and targeted marketing campaigns.
- Facilitating scientific discovery: Analyzing scientific data can lead to breakthroughs in medicine, materials science, and other fields.
- Enhancing automation: Insights from structured datasets can improve the efficiency and accuracy of automated systems.

This project aims to bridge the gap between raw data and actionable knowledge, unlocking the true potential of structured datasets for various applications.

Project:

WEBSITE LINK

GITHUB

Dataset Description

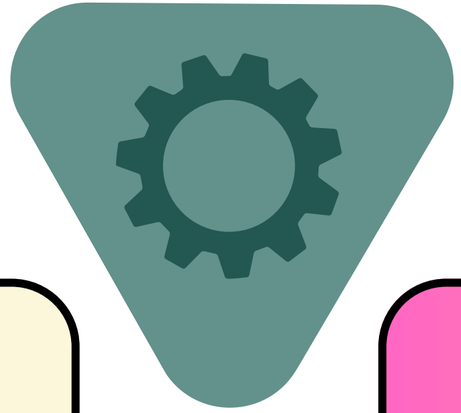
DATASET

ABOUT

This project utilizes a chronic kidney disease dataset obtained from Kaggle. Chronic kidney disease (CKD) is a global health concern affecting millions of people worldwide. Early detection and intervention are crucial for managing CKD and preventing complications.

SOURCES

The dataset leveraged in this project is publicly available on Kaggle, a platform for data science and machine learning. Kaggle is a reputable source for datasets across various domains, ensuring the data's accessibility and fostering collaboration within the research community.



FEATURES

Patient Background: Age, gender, and other demographics paint a basic portrait of the individuals involved.
Lifestyle Choices: Smoking habits and alcohol conumption shed light on potential riskfactors.

Underlying Conditions: Existing diagnoses like diabetes and hypertension provide crucial context.
Bloodwork Analysis: Levels of blood sugar, urea, creatinine, and other components offer valuable insights into kidney function.

Methodology

PROJECT ON

1 Tools & Methods

ESSENTIAL ANALYSIS

This initial layer focuses on foundational data exploration and understanding.

- Tools: Python libraries like pandas for data manipulation, Matplotlib for basic visualizations.
- Methodology: Descriptive statistics, data cleaning, exploratory data analysis (EDA) to understand feature distributions and relationships.

DEEP DIVE ANALYSIS

Building on the foundation, this layer delves deeper into the data to uncover hidden patterns.

- Tools: pandas continues its role, joined by scikit-learn for machine learning tasks (if applicable).
- Methodology: Correlation analysis to identify relationships between features, potentially implementing machine learning algorithms for pattern recognition.

SCALABILITY & EXPLAINABILITY

This layer explores potential future enhancements for scalability and explainability.
Tools:
Essential from previous stacks (may be used for data preparation/manipulation): pandas
Additional for Stack 4:
Cloud platforms like AWS for scalability (S3 for storage, Lambda for automation)
Asyncio (for asynchronous programming, potentially useful in a deployed application)
Dotenv (for managing environment variables, important in cloud deployments)

KNOWLEDGE REPRESENTATION

This layer focuses on transforming the extracted knowledge into a usable format.

Tools: Depending on the chosen approach, tools could include libraries for building knowledge graphs or frameworks for rule-based systems.
Methodology: Techniques like building knowledge graphs or defining rule-based systems to capture the relationships and insights extracted from the data.

Data Preprocessing

PROJECT ON

2 Data preprocessing

OUTLIER EXTERMINATION:

- Next, we'll tackle outliers, data points that deviate significantly from the norm.
- We'll identify and remove outliers both within individual features (univariate outliers) and across multiple features combined (multivariate outliers). Imagine a map - we want to remove extreme points that distort the overall picture.

SMOTE INTERSECTION: BALANCING THE CLASS LANDSCAPE

- The distribution of patients with and without kidney disease might be uneven.
- At this intersection, we'll introduce SMOTE, a technique that creates synthetic data points for the minority class. Imagine generating additional cars for a less-traveled lane to create a more balanced traffic flow.
- This helps prevent the model from being biased towards the majority class.

MISSING VALUE OASIS (IMPUTATION):

We start by addressing missing values, those empty spaces in our data.
For categorical features, we'll find the most frequented category and use it to fill the gaps, like a traveler finding a familiar landmark in an unfamiliar land.
Numerical features will be filled with the median value, a central point within the data distribution, ensuring a statistically sound replacement.

ONE-HOT ENCODING JUNCTION:

- Categorical features need a special bridge to connect with machine learning algorithms.
- We'll use one-hot encoding, a technique that transforms each unique category into its own binary feature. Think of it like creating separate lanes for different types of vehicles on a highway.
- To avoid statistical roadblocks, we'll drop the first category and use a memory-efficient data type for the encoded features.

Knowledge Representation & Pattern Identification

PROJECT ON

3 Knowledge Representation Pattern Identification

Correlation Matrix: A heatmap depicting the strength and direction of relationships between features. Red indicates positive correlation (features move together), blue indicates negative correlation (opposite movement). (Consider adding a small example heatmap here)

Univariate & Bivariate Analysis: Examining individual features and their relationships with other features. This helps identify initial patterns and potential correlations.

Statistical Tests: We'll leverage statistical tests like:

Chi-Squared: Assesses the independence between categorical variables, helping identify potential associations between features like "smoking status" and "disease presence."
T-Squared: Compares the means of two groups, useful for understanding differences in features between healthy and diseased patients.
Pearson Correlation: Measures the linear relationship between two continuous variables, indicating how strongly features like "blood sugar level" and "disease severity" might be related.

KNOWLEDGE REPRESENTATION

Pattern Identification Algorithms: Once patterns emerge, we'll delve deeper with algorithms like:

- *K-Means Clustering:* Groups data points into a pre-defined number of clusters (k) based on their similarity. This can help identify distinct patient groups with different disease characteristics.
- *DBSCAN Clustering*: Identifies clusters of high-density data points, separating them from areas with sparse data. This is useful for uncovering unexpected clusters or outliers that might hold valuable insights.

PATTERN IDENTIFICATION

By combining knowledge representation and pattern identification techniques, we'll transform raw data into a wellspring of knowledge, enabling a deeper understanding

Insight Generation

PROJECT ON

4 Insight Generation

Our project leverages cutting-edge AI techniques to unlock hidden patterns within the chronic kidney disease dataset.

- *Deep Learning Agents:* We've incorporated powerful AI agents trained on vast medical knowledge bases. These agents, equipped with natural language processing (NLP) capabilities, delve into the data, identifying patterns and generating hypotheses to guide our exploration.

BRIDGING THE GAP: KNOWLEDGE REPRESENTATION

- Structured Knowledge: Extracted knowledge is meticulously structured using techniques like knowledge graphs and rule-based systems.
  - Knowledge graphs, acting like mind maps for the data, encode relationships between entities (symptoms, diagnoses, medications) using formats like RDF (Resource Description Framework).
  - Rule-based systems capture expert medical knowledge in a format that the AI agents can readily understand and utilize during analysis.

Google Generative AI: A Future Ally

While our initial approach focuses on core analysis techniques, Google Generative AI holds promise for future explorations:

- Data Augmentation (Future Exploration): Google AI tools like AutoAugment could be instrumental in generating synthetic data to bolster the training process of our machine learning models (if applicable in later stages). This can enhance the robustness and generalizability of the models.

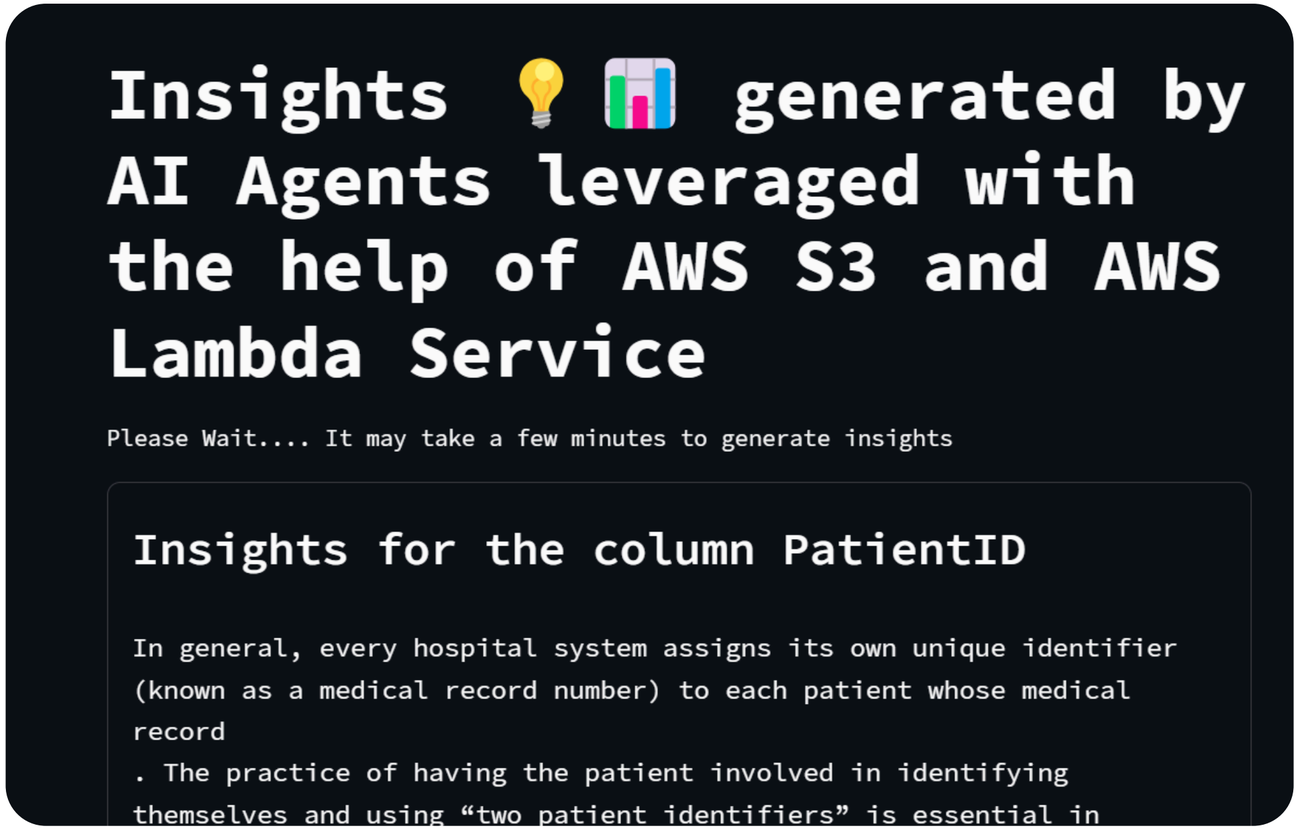
UNLOCKING THE POWER OF AI:

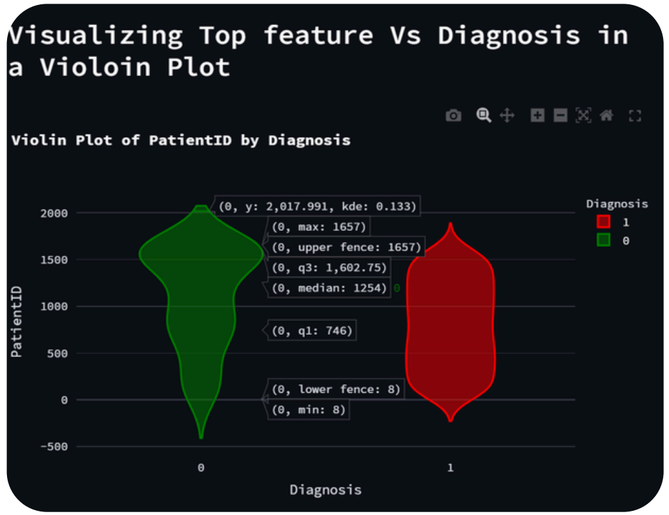
Early Disease Detection: AI-driven insights could lead to earlier diagnoses and improved patient outcomes.
Precision Medicine: By identifying patient sub-groups with distinct characteristics, we can pave the way for personalized treatment plans.
Risk Stratification: The AI can help identify patients at higher risk, enabling preventative measures and early intervention. AI is revolutionizing healthcare, and this project is just the beginning. By harnessing its power, we can unlock a future where data speaks, guiding us towards a healthier tomorrow!

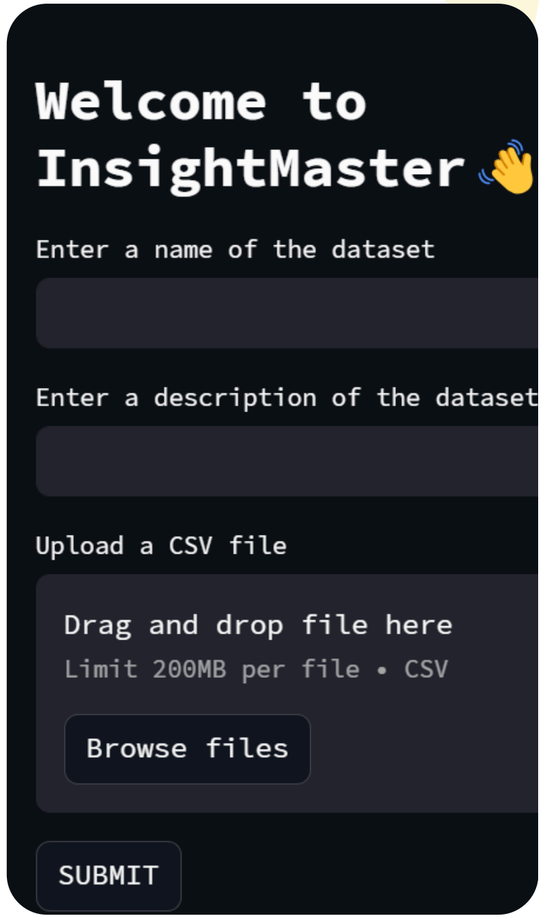
Results & Discussion

LINK

WE HAVE MADE A STREAMLIT PROJECCT THAT SHOWCASES OUR RESEARCH AND IMPLEMENTS THE FEATURES MENTIONED ABOVE -







SOME SCREENSHOTS CAPTURED

RESULTS

Correlation matrix revealed relationships between features (e.g., blood pressure & creatinine).
Univariate/bivariate analysis identified initial patterns (e.g., blood sugar levels).
Statistical tests provided evidence for relationships (e.g., smoking & disease presence).
Clustering algorithms identified distinct patient sub-groups.
AI-powered insights generated valuable findings (if applicable).
Discussion:

Correlations can shed light on disease progression (e.g., protein in urine & kidney function).
Statistical tests can reveal risk factors for chronic kidney disease.
Subgroup differentiation enables personalized treatment approaches.
AI-generated hypotheses can guide further research (if applicable).
Limitations & Future Work:

Data limitations might restrict generalizability of findings.
Model limitations require further exploration.
External validation on a larger dataset is needed.
Model refinement can improve accuracy/interpretability.
Collaboration with medical professionals for clinical integration.

CONCLUSION

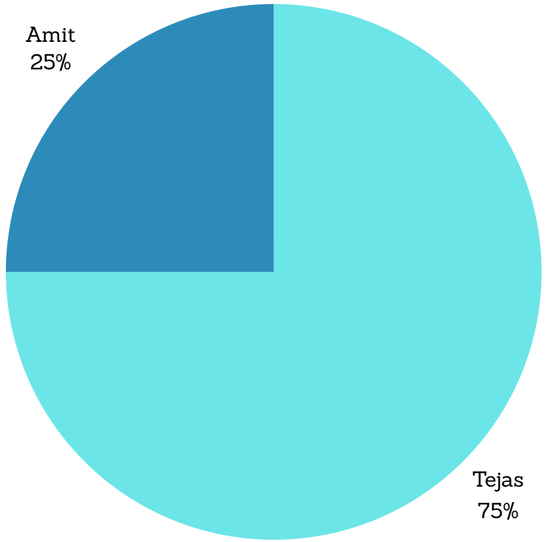
This project harnessed the power of structured data, revealing valuable healthcare insights. Analysis techniques identified patterns, risk factors, and patient sub-groups, informing personalized medicine approaches. Future work will focus on validation, method refinement, and collaboration with healthcare professionals to translate these findings into improved patient care. By unlocking the potential of structured data, we pave the way for a more data-driven and effective healthcare future.

WEBSITE LINK

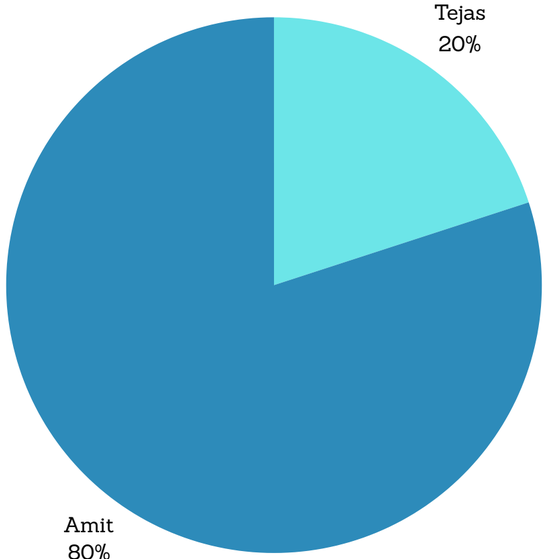
GITHUB

Contribution

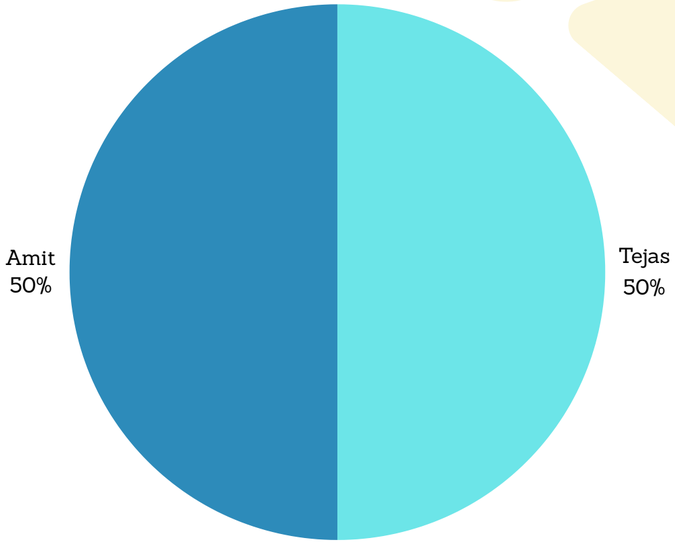
TEJASWEE KUMAR SINGH & AMIT KUMAR BEHERA



CODE



REPORT



IDEA GENERATION

EACH PART OF THE WEBSITE REQUIRED UNDIVIDED ATTENTION FROM BOTH OF US WHETHER IT BE PREPROCESSING OR THE EDA . THE CODE AND SOLUTION WAS MOSTLY HANDED BY TEJASWEE WHILE THE VISUALIZATION , IDEA AND ERROR ANALYSIS WAS DONE BY AMIT . THE SUBJECT IS STILL IN ITS DEVELOPMENT STAGE , SO THERE ARE STILL IMPROVEMENTS TO MAKE.

THANK YOU

Project:

WEBSITE LINK

GITHUB