DOC-AI

A Medical AI-Chatbot

- By: How I Met Your Data

Our team

Vicker Ivy
Presentation

Daisy Kerubo Thomas
Jupyter notebook

Victor Ongaki
Deployment master

Felix Musau
Scram master

Rose Matoke
Github managment

Introduction

In today's fast-paced hospital environments, **patients often face long wait times, limited access to medical professionals, and confusing online health information**. To help bridge this gap, our team developed a conversational **medical chatbot using the Disease and Symptoms 2023 dataset by Mendeley data.**

This will be done by implementing **machine learning and natural language processing.**

Problem Statement

01

What?
There is a pressing need for an **intelligent, secure, and accessible medical chatbot t**hat can provide reliable health guidance, reduce the workload on healthcare professionals, and improve patient engagement while overcoming challenges related to accuracy, empathy, data protection, and connectivity.

02

Why?
The healthcare sector faces **communication gaps between patients and doctors** due to high workloads, limited consultation time, and restricted access to reliable medical advice. Patients often experience delays in receiving care, while doctors handle repetitive inquiries that reduce efficiency.

Main objective

To develop and implement a conversational medical chatbot system within a hospital setting that enhances healthcare guidance for patients and categorizing patient diseases based on their symptoms.

specific objectives

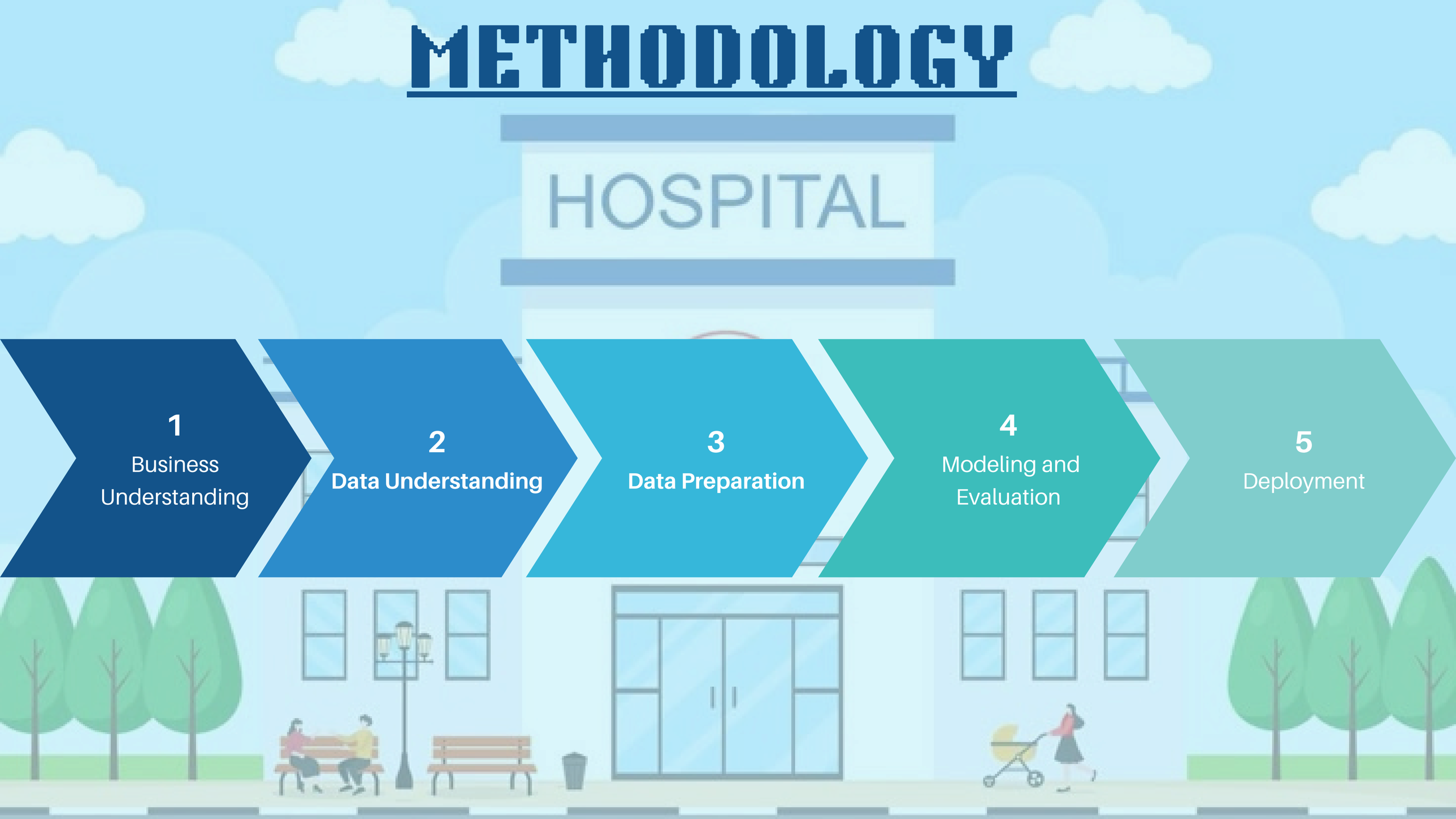
01 To provide 24/7 automated medical support to help patients get quick answers to patient diseases based on symptoms even outside normal working hours.

02 To reduce the workload of nurses, doctors' and receptionists by handling routine tasks and FAQs

03 To enhance patient experience and engagement by delivering easy-to-understand responses that improve satisfaction within the hospital ecosystem.

04 To efficiently collect patient information and symptoms through a user-friendly interface, potentially reducing wait times and improving patient experience.

METHODOLOGY



1
Business Understanding

2
Data Understanding

3
Data Preparation

4
Modeling and Evaluation

5
Deployment

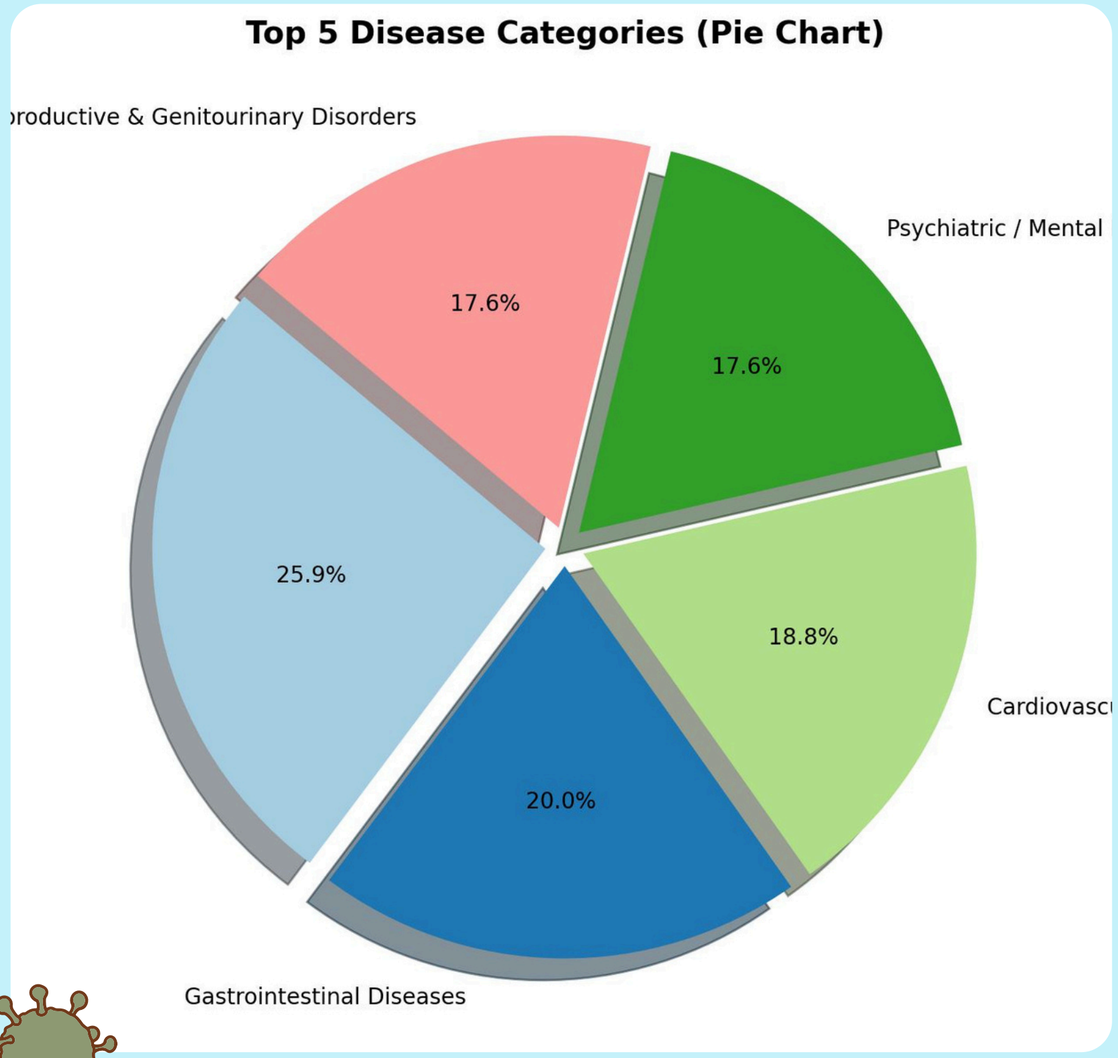
Data Understanding

Data: Disease and Symptoms Dataset(2023) from Mendeley data

- **Rows:** 246,945k (diseases)
- **Columns:** 378 (symptoms)

citation:
Stark, Bran (2025), “Disease and symptoms dataset 2023”, Mendeley Data, V1, doi: 10.17632/2cxccsxydc.1

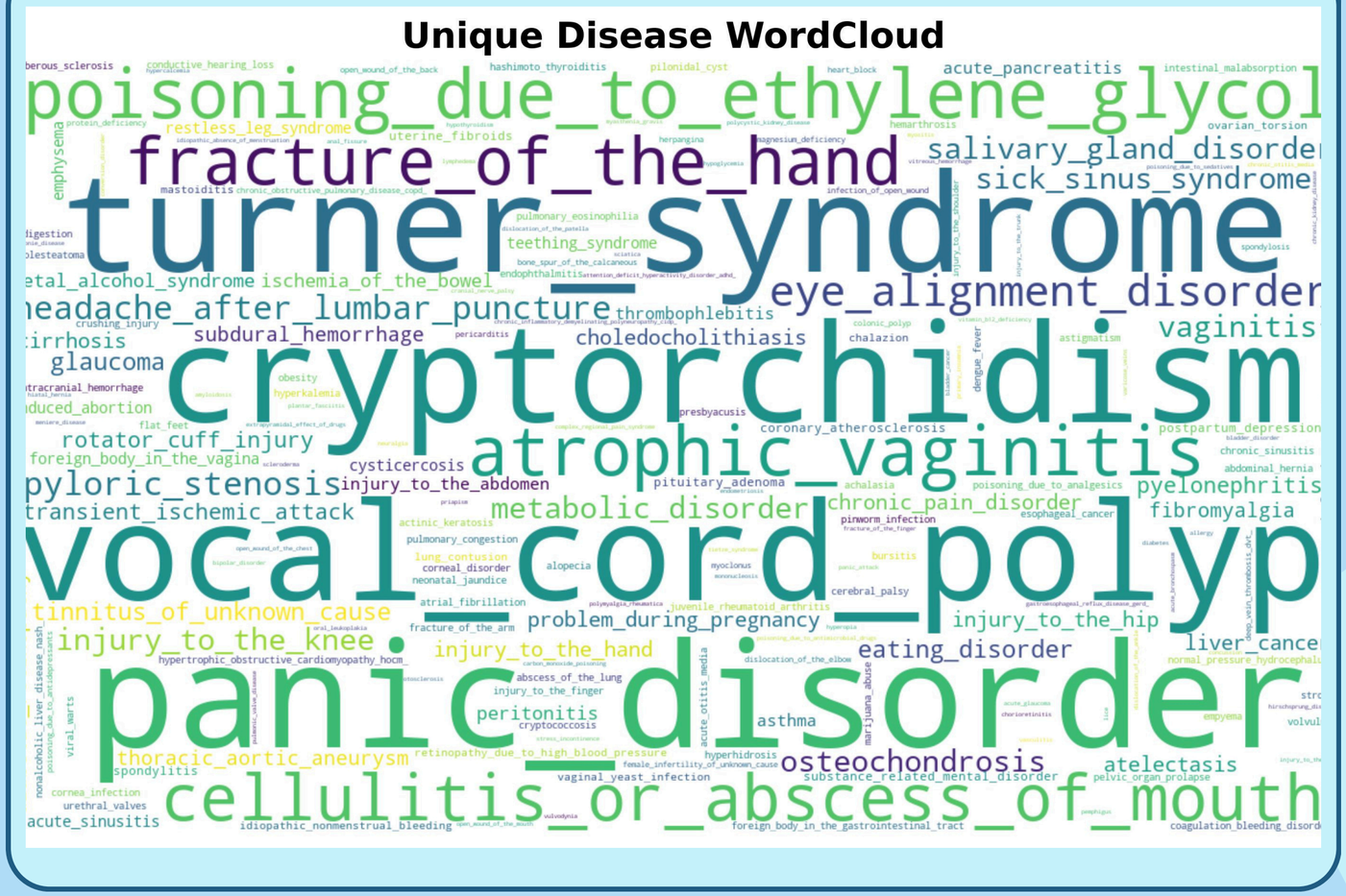
DATA ANALYSIS



DOMINANT DISEASE CATEGORY

- **Infectious Diseases (25.9%)** form the largest category, indicating their wide diversity and global prevalence.
- **Gastrointestinal Diseases (20.0%)** rank second, showing the significant impact of digestive system disorders such as ulcers and liver disease.
- **Overall, the top five categories** show that diseases affecting essential body systems i**mmunity, digestion, circulation, mental health, and reproduction** dominate the global disease landscape.

MOST FREQUENT DISEASES



- **Prominent Diseases (largest words)**

The biggest terms like **panic disorder**,**vocal cord polyp**,**cryptorchidism**,**turner syndrome** and **poisoning due to ethylene glycol** appear most clearly. In a WordCloud, size represents frequency or importance, so these diseases are likely the most common or most emphasized entries in the dataset.

- **Rare or specialized conditions**

Smaller words such as **thoracic aortic aneurysm**, **subdural hemorrhage**, or **choledocholithiasis** appear much less frequently these could represent specialized or rare cases in our data.

MODELLING

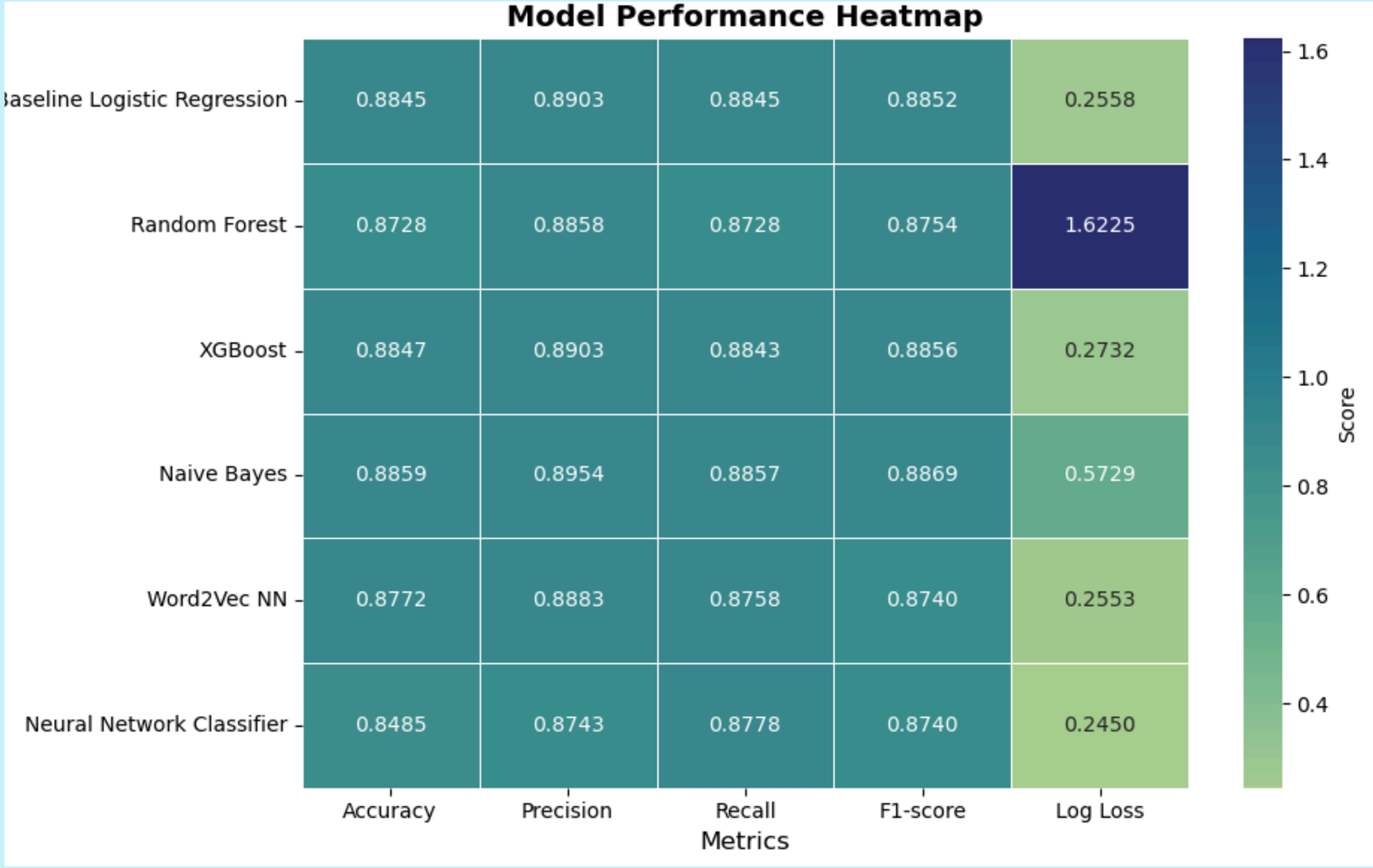
To build an effective disease classification system, **we applied several machine learning models after filtering the dataset to include diseases with more than 800 occurrences.**

The data was balanced using **SMOTE and split into training, validation, and test sets.**

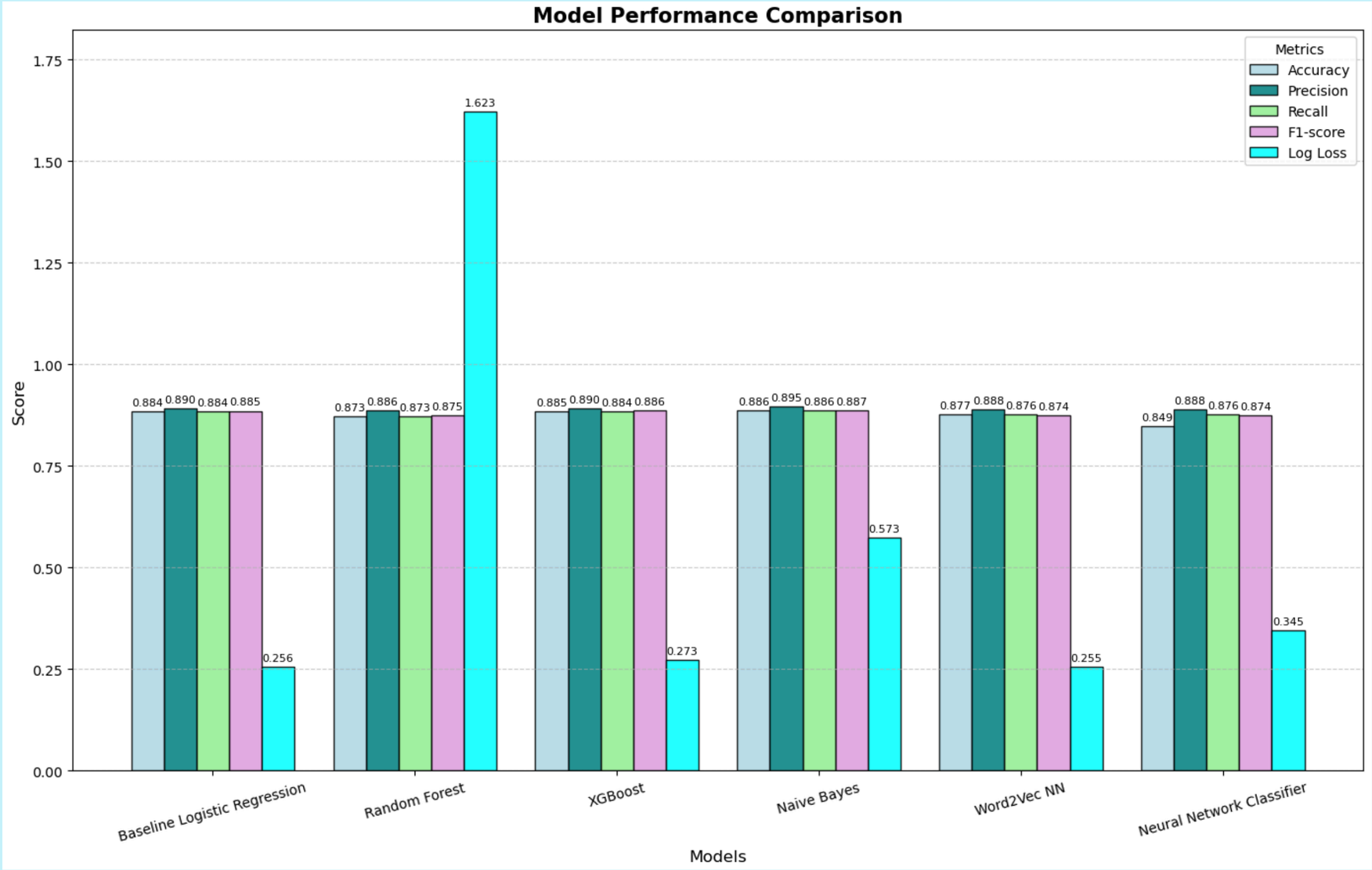
We made use of **classifier models such as Logistic Regression(baseline model), Random Forest, Naive Bayes, Word2Vec NN, Neural Network Classifier and XG Boost.**

Overall, the models demonstrated high reliability and generalization, with **XG Boost and the Neural Network emerging as the best-performing models.**

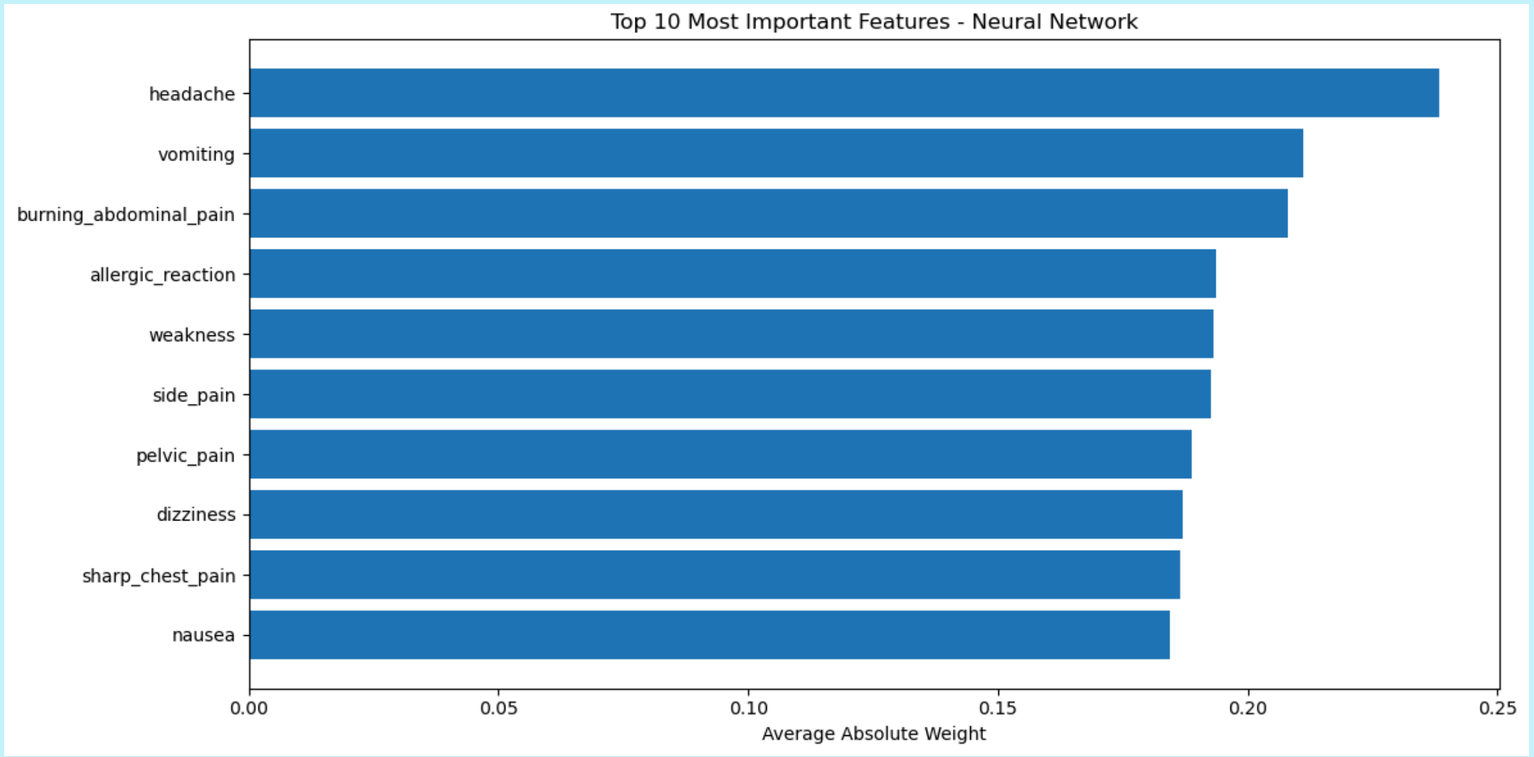
Model Performance Heatmap



Model Performance Comparison



Top 10 Most Important Features – Neural Network



Headache, vomiting, and burning abdominal pain rank as the first three in terms of feature importance.

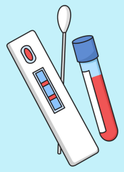
Model Evaluation

- The neural network achieves **Accuracy = 87.72%, Precision = 88.83%, Recall = 87.58%, and F1 = 87.40%, log _loss** = **25.53%** all within a tight range.

- In comparison, models like Random Forest or XG Boost often show slightly higher training accuracy but drop more on test data (signs of mild overfitting).

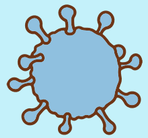
- The **neural network generalizes well to unseen data**, which is crucial for real-world deployment. While overall Accuracy remains high(0.8772),it is the **combination of high Recall and low Log Loss** that makes this neural network the **most dependable and deployment ready model for real-world use**.

Conclusions



We built a smart system that can **suggest possible diseases based on symptoms shared by patients.**

Our best model **correctly predicted diseases about 87% of the time**, showing strong reliability.



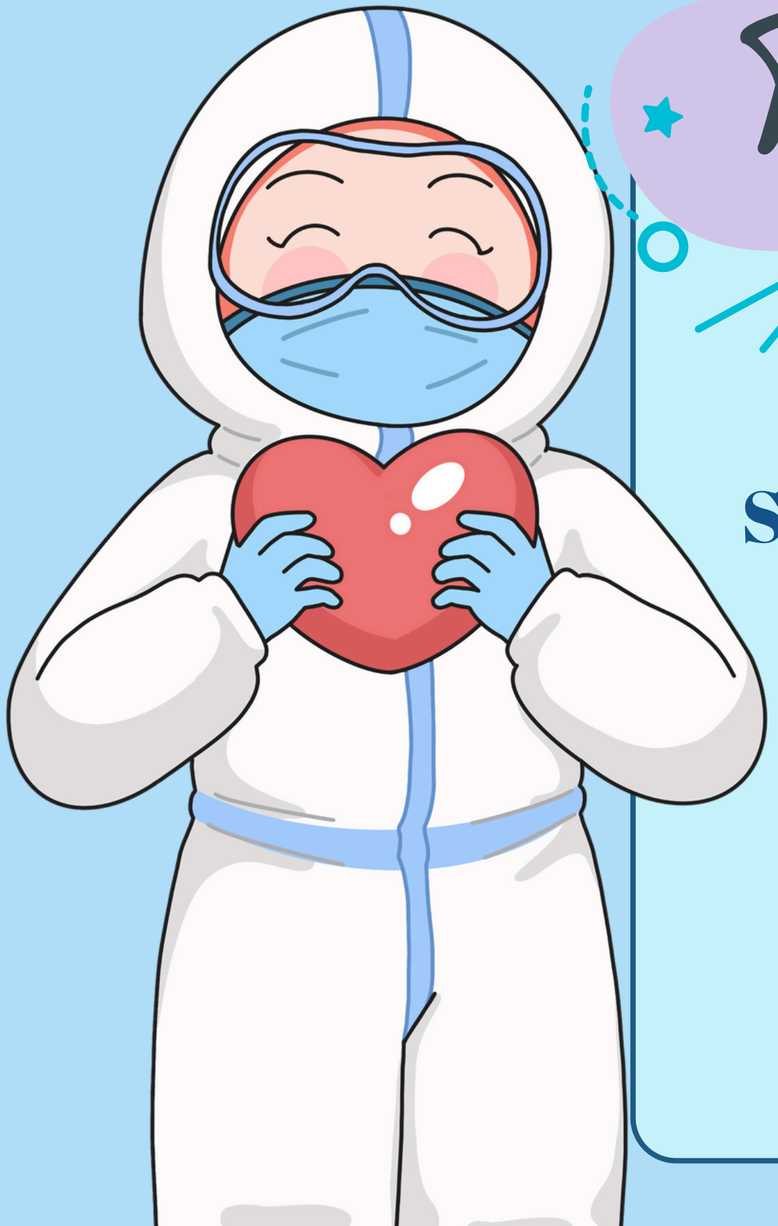
- The **most frequently predicted diseases were cystitis, and nose disorders** showing a wide range of health conditions in our data.
- Some symptoms, like **fever and cough, appeared across many diseases**, making them key indicators in diagnosis.

I**nfectious diseases were the most common overall, followed by stomach and heart-related conditions** highlighting how lifestyle and aging affect health

NEXT STEPS

- **Connect to Trusted Databases:** Link with WHO, MedlinePlus, and similar sources for accurate, up-to-date medical info.
- **Build cross platform App:** Make the app easily accessible using popular platforms such as google play store on any device.
- **Use Reinforcement Learning:** Enable the chatbot to learn from user feedback and improve over time.
- **Add Adaptive Conversations:** Use follow-up questions when confidence is low to boost diagnostic accuracy.
- **Promote Preventive Healthcare:** Support integrated, data-driven strategies for early diagnosis and mental health.
- **Ensure Privacy & Ethics:** Comply with data protection laws and ethical standards to safeguard user trust.

“I’d love to hear your thoughts or questions!”



Thank you

Scan the QR below to access our web app

SCAN ME