Demystifying Disease Risk: A Visual predictor

Making ML Predictions Transparent for Symptom-Based Health Insights

Project Team: Alex Ivanoff, Noah Schmidt, Pojanart Foote, Stephanie Fritz

The “Black Box” Problem in Health AI

The Power of AI in Healthcare

- Machine Learning (ML) is **transforming** healthcare.
- It’s **used for everything** from predicting disease outbreaks to aiding in diagnoses and personalizing treatment plans.

The Challenge

- Many powerful ML models are **complex** inherently opaque.
- It’s **hard to understand** *how* they arrive at their predictions
- They **act like a “Black box”** – inputs go in, predictions come out, but the internal logic is hidden.

Impact in Healthcare

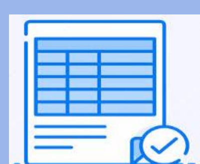
- Leads to distrust and misunderstanding.
- Limits effective action based on AI insights.

OUR SOLUTION: Apply Data Visualization & Analytics to make AI transparent

Our Solution: The Visual Predictor (High-Level)

User – Friendly Symptoms Input

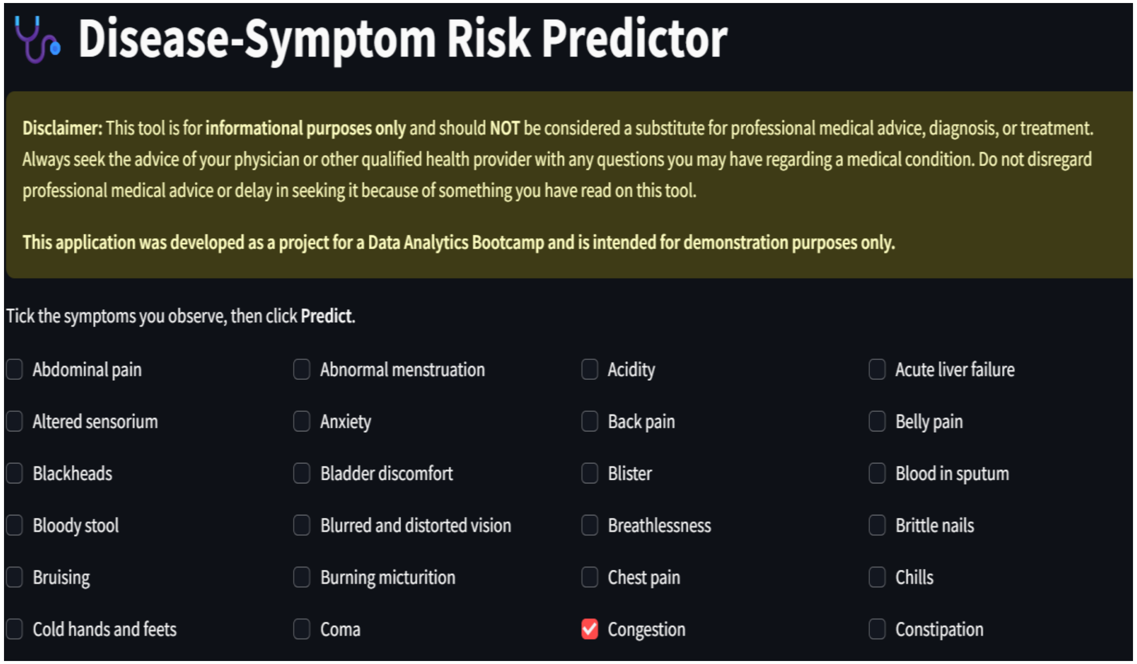
Our ML – Powered Risk Assessment

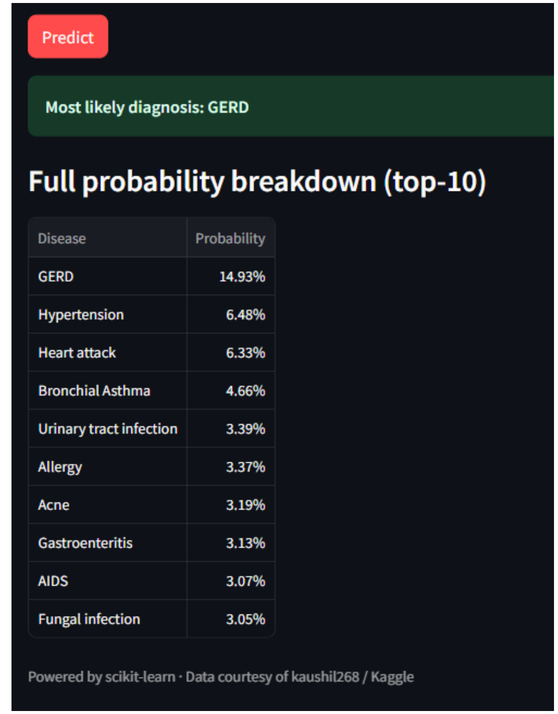


Transparent Output: Beyond a Single Prediction

✓ What: Probability Breakdown (Top-10)

What We Delivered





- **Intuitive Symptom Input**: User-friendly web interface with symptom selection
- **ML Model for Risk Prediction:** Identifies potential disease based on selected symptoms
- **Core Transparency Features**
    - Full Probability Breakdown (Top 10): Shows model’s confidence across top diseases

Core Technologies & Libraries

Data & ML Core

scikit learn

DataFrame Pandas

Joblib

SQLite

Web Delivery & UI Interface

Streamlit

Visuals & Collaboration

matplotlib

GitHub

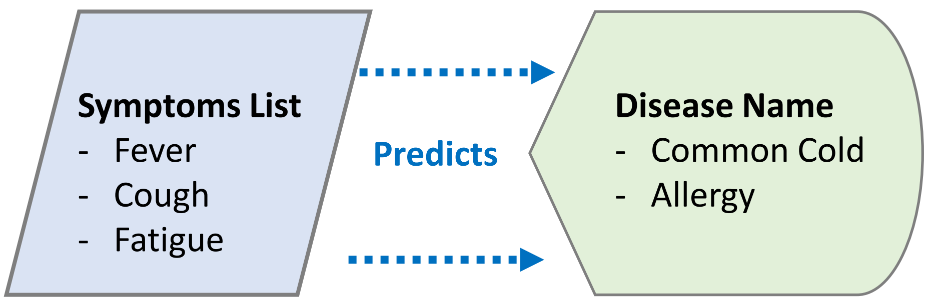
Our Data: Symptoms to Disease

kaggle

✓ **Data Source:** Kaggle's Disease-Symptom Description Dataset *(Clear symptom-to-disease mappings needed for our predictive model)*

Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4
Diseases that may be present	the symptoms experienced during the disease	the symptoms experienced during the disease	the symptoms experienced during the disease	the symptoms experienced during the disease
41 unique values	vomiting 17% fatigue 14% Other (3408) 69%	vomiting 18% fatigue 8% Other (3648) 74%	fatigue 15% high_fever 7% Other (3870) 79%	high_fever 8% [null] 7% Other (4194) 85%
Fungal infection	itching	skin_rash	nodal_skin_eruptions	
Fungal infection	itching	skin_rash	nodal_skin_eruptions	dischromic _patches
Allergy	continuous_sneezing	shivering	chills	watering_from_eyes
Allergy	shivering	chills	watering_from_eyes	
Allergy	continuous_sneezing	chills	watering_from_eyes	
Allergy	continuous_sneezing	shivering	watering_from_eyes	
Allergy	continuous_sneezing	shivering	chills	
Allergy	shivering	chills	watering_from_eyes	

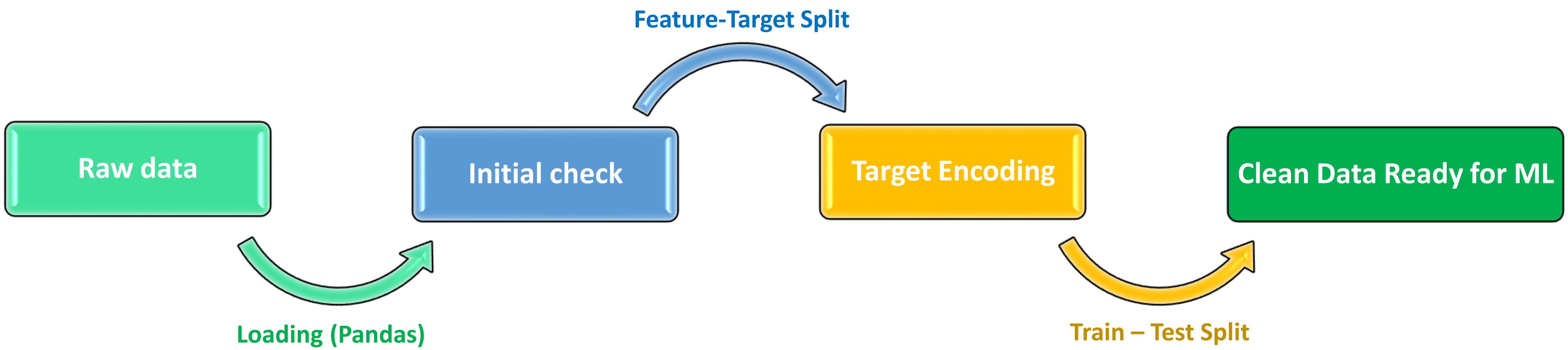
✓ **Structure:** Symptoms directly mapped to disease prognoses. *(Initially CSVs, pre-processed into an SQLite database for efficient access. Allowing us to directly use symptom presence as features)*



✓ **Relevance:**
Ideal for our symptom-based disease prediction model

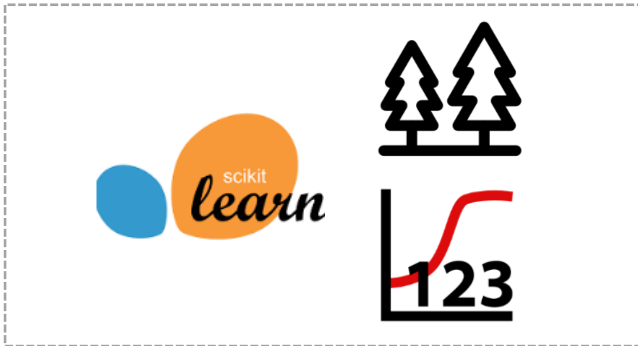
✓ **Format:**
Pre-cleaned and stored in an SQLite database

Preparing Data for Prediction



- **Loading & Initial Checks:** From SQLite database into Pandas DataFrame.
- **Feature – Target Split:** Separated symptoms (features) from disease (target)
- **Target Encoding:** Converted text disease names to numerical labels (and back for output)
- **Train-test Split:** Divided data for fair model evaluation and to prevent overfitting.

Choosing Our Prediction Engine



Choose Model

Problem Type: Supervised Classification.

Algorithms Explored (Scikit-learn):

- ✓ **Logistic Regression:** Efficient, probabilistic.
- ✓ **Random Forest:** Robust, high accuracy.

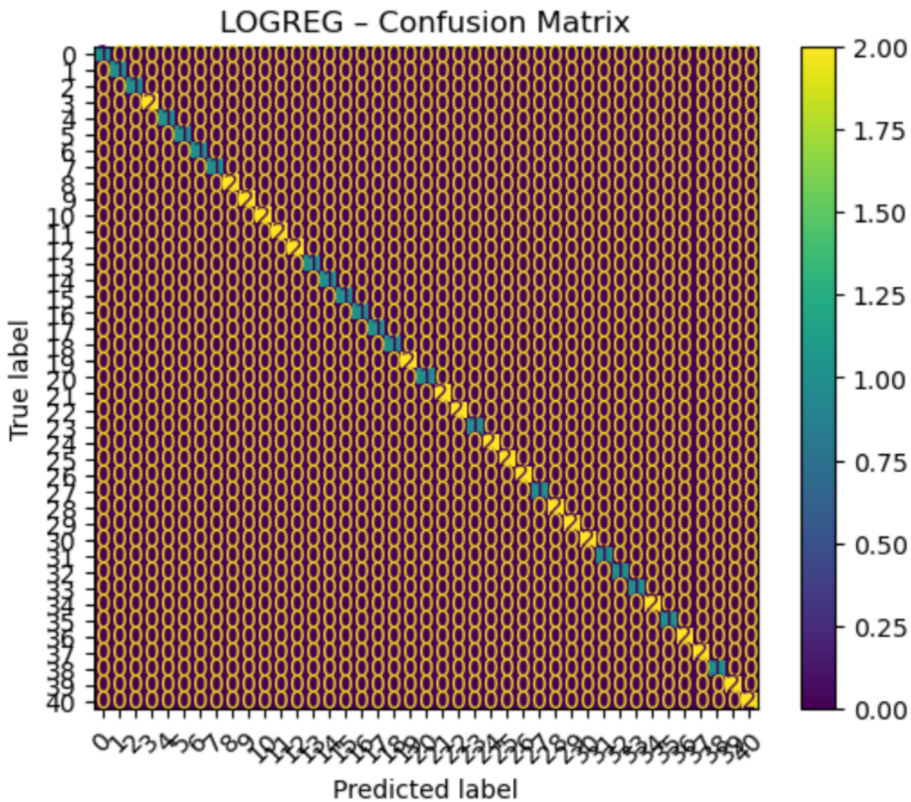
Selection

Chose the best performer based on predictive accuracy and suitability for probability output.

Training Data → ML Model → Prediction

Training: Model trained on our prepared data.

Evaluating Our Model's Effectiveness



- **Assessing Reliability:** Crucial step to ensure trustworthy predictions.

- **Key Metrics:**
  - **Accuracy:**
    ```
    logreg  accuracy: 1.000
    rf      accuracy: 1.000
    ```
  - **Confusion Matrix:** Provides detailed breakdown of correct vs. incorrect classifications.
- **Limitations:**
  - Only predicts 42 diseases
  - Only takes symptoms as input and does not consider other medical history information

OVERALL: Our model demonstrates strong predictive performance.

Our Virtual Predictor in Action!

Live Demo

Future Enhancements

- ➢ Enhanced Symptom Input (e.g., search, NLP)
- ➢ More Advanced Visual Explanations

- ➢ More back-end data
- ➢ Potential to predict more diseases

- ➢ Ability to input medical history and have it impact results

  (more data would be necessary for that).
- ➢ User Feedback & Model Refinement Loop

- ➢ Public Deployment

Overcoming Hurdles & Key Takeaways

Challenges Faced:

- Managing scope in a 2 week-sprint
- Integrating ML backend with Streamlit frontend
- Designing effective demystification visuals

Key Learnings:

- Practical application of the full ML workflow
- Importance of user-centric AI design for transparency.


THANK YOU!