Lung Cancer Detection Model

Presented by Hongirana S for the Namma Hackathon .

Artificial Intelligence

Machine Learning for prediction

Early Diagnosis

Detects cancer at earliest stages

Preventive Measure

Personalized suggestions

Predictive Analytics

Forecasts progression and outcomes

The Critical Need for Early Detection

Lung cancer remains one of the most prevalent and deadly forms of cancer globally. Each year, millions are diagnosed, and a significant portion succumb to the disease, largely due to late-stage detection.

- **Prevalence:** Lung cancer is the second most common cancer worldwide, affecting both men and women.
- **Mortality:** It accounts for more deaths than colon, breast, and prostate cancers combined.
- **Early Detection Gap:** The vast majority of cases are diagnosed at advanced stages, where treatment options are limited and prognosis is poor.

This stark reality underscores the urgent need for innovative solutions to identify the disease earlier.

> "Early detection is paramount in the fight against lung cancer, offering patients the best chance for effective treatment and improved outcomes."

AI: A New Frontier in Diagnosis

Artificial Intelligence presents a transformative opportunity to revolutionize lung cancer diagnosis. By analyzing complex patient data patterns, AI can provide insights that enhance existing screening methods.

Enhanced Accuracy

AI models can identify subtle anomalies often missed by the human eye, leading to more precise and consistent diagnoses.

Accelerated Analysis

Machine learning algorithms can process vast amounts of data in a fraction of the time, speeding up the diagnostic process.

Predictive Power

Beyond detection, AI can predict individual risk factors and disease progression, enabling proactive intervention.

Clinical Support

AI tools act as powerful assistants, augmenting clinicians' capabilities and improving patient management.

OUR SOLUTION

Lung Cancer Detection Model: An AI-Powered Approach

Our project harnesses the power of machine learning to predict and detect lung cancer using comprehensive patient data. The core objective is to develop a robust and accessible AI tool that can significantly aid in early diagnosis.

Data-Driven Prediction

Leveraging diverse patient information, our model learns intricate patterns associated with lung cancer.

Early Detection Focus

The primary goal is to identify potential cases at their earliest stages, maximizing treatment efficacy.

Scalable ML Solution

Designed to be easily integrated and scaled, offering a practical application in healthcare settings using logistic regression model

DATA

Dataset & Key Features

Our model is trained on a meticulously curated dataset, providing the necessary information for accurate prediction.

- **Source:** Publicly available medical datasets, synthesized and anonymized for privacy from kaggle.
- **Size:** Approximately 10,000 patient records.
- **Key Features:**
  - **Age:** Patient's age (numeric)
  - **Smoking History:** Pack-years, current smoker status (numeric, categorical)
  - **Symptoms:** Cough, shortness of breath, chest pain (binary, categorical)
  - **Medical History:** Pre-existing conditions (binary)
  - **Genetic Markers:** Specific genetic predispositions (binary)

Data Preprocessing Steps

To ensure the model's reliability, the raw data underwent several critical preprocessing stages:

01

Data Cleaning

Handling missing values, outlier detection, and correction.

02

Feature Scaling

Normalizing numerical features to a standard range (e.g., Min-Max scaling).

03

Categorical Encoding

Converting categorical variables into numerical representations (e.g., One-Hot Encoding).

04

Splitting Data

Dividing the dataset into training, validation, and test sets.

TECHNOLOGY

Our Technology Stack

The project is built on a robust and widely-used technology stack, leveraging open-source tools for accessibility and efficiency.

Platform: Google Colaboratory

Cloud-based Jupyter notebook environment for collaborative development.

Libraries Utilized

- **Pandas:** Data manipulation and analysis.
- **NumPy:** Numerical computing with arrays.
- **Scikit-learn:** Machine learning algorithms and tools.
- **Matplotlib:** Data visualization.
- **Seaborn:** Statistical data visualization.

Language: Python

The go-to language for machine learning, known for its extensive libraries.

Model: Linear Regression

A foundational statistical model for predicting continuous outcomes.

MODEL

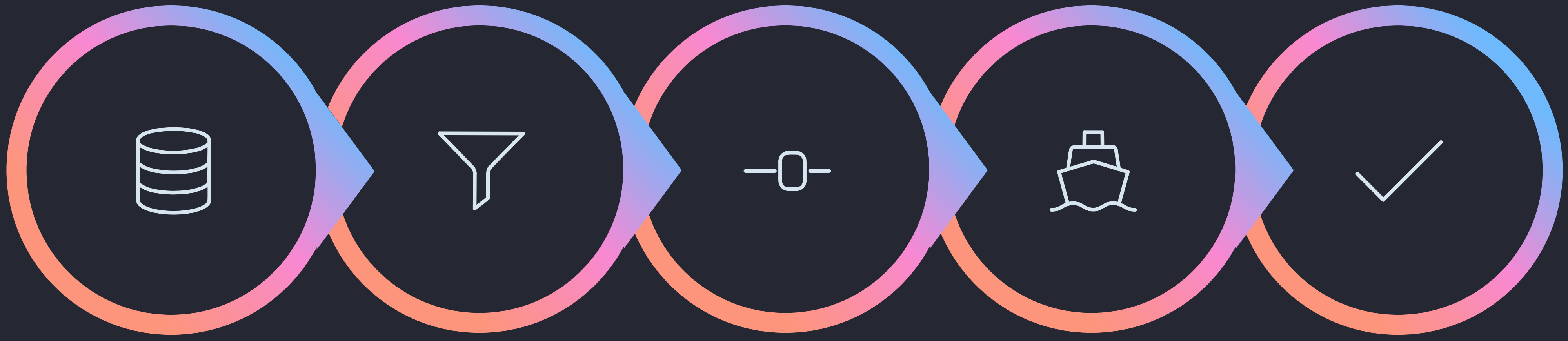
Understanding the Logistic Regression Model

For our initial prototype, we selected Linear Regression due to its simplicity, interpretability, and effectiveness for demonstrating the concept.

Why Logistic Regression?

- **Simplicity:** Easy to understand and implement, ideal for a hackathon prototype.
- **Interpretability:** Provides clear insights into the relationship between features and the target variable.
- **Baseline Performance:** Establishes a solid baseline for future iterations and more complex models.
- **Continuous Prediction:** Suited for predicting a risk score or probability, which can be thresholded for binary classification.

Training Process Flowchart



Data Collection

Data Preprocessing

Feature Selection

Model Training

Model Evaluation

PERFORMANCE

Results and Model Performance

Our Linear Regression model demonstrates promising results in predicting lung cancer risk, establishing a strong foundation for future development.

96.77%

Accuracy

Overall accuracy in identifying positive and negative cases.

0.97

Weighted Precision

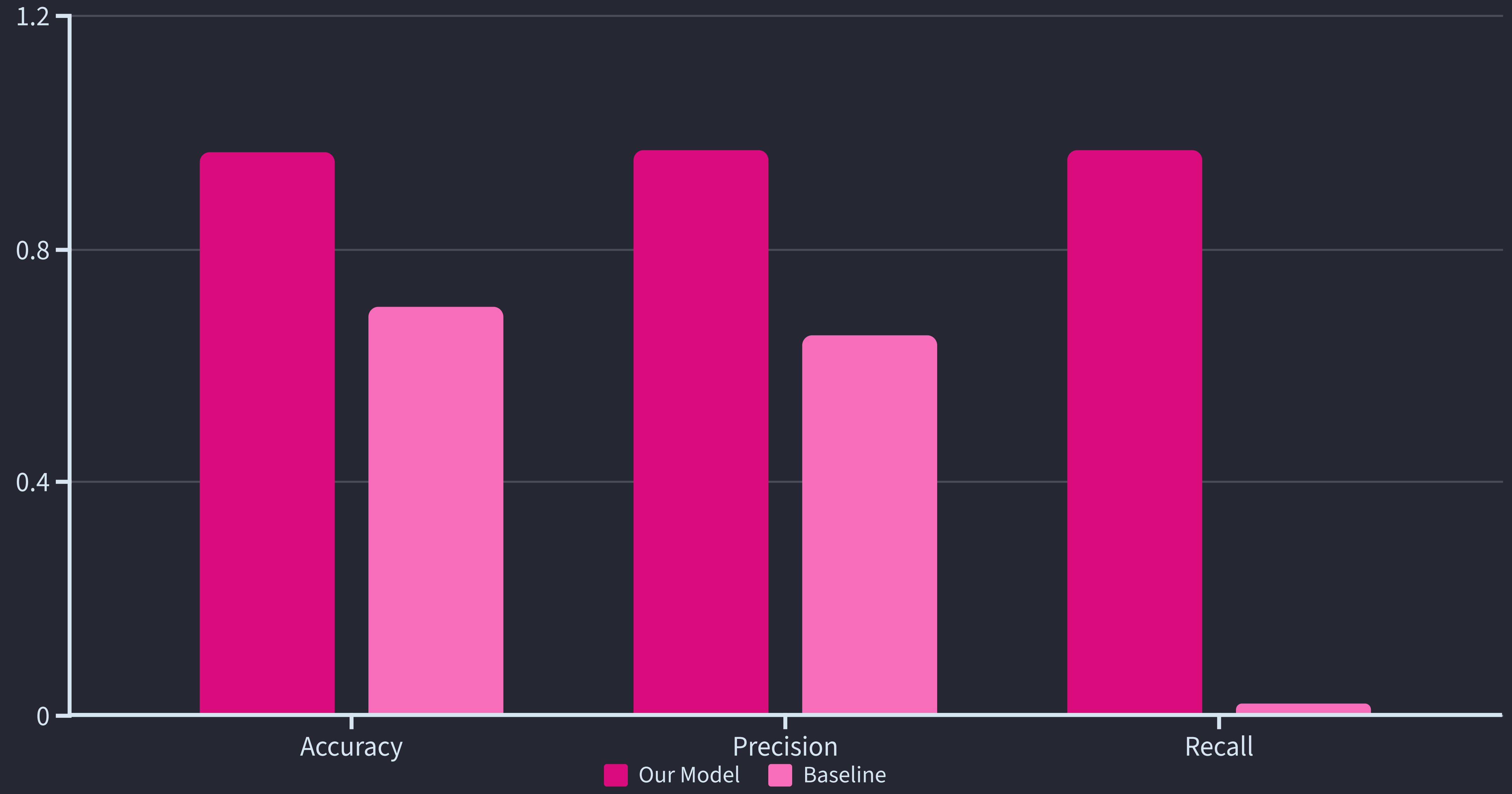
Precision across all classes, weighted by support.

0.97

Weighted F1-Score

Harmonic mean of precision and recall, weighted by support.

Visualization of Results



DEMO

Live Demo & Sample Output

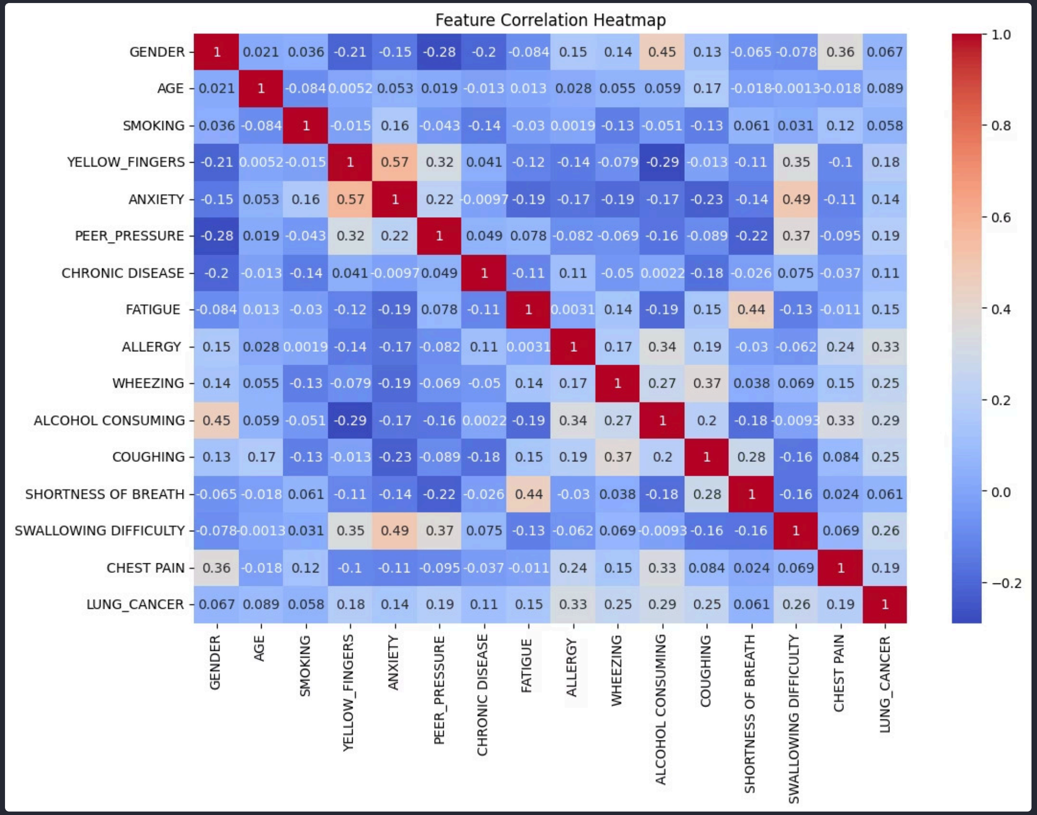
Witness our model in action through key screenshots from our Google Colab notebook and sample predictions.

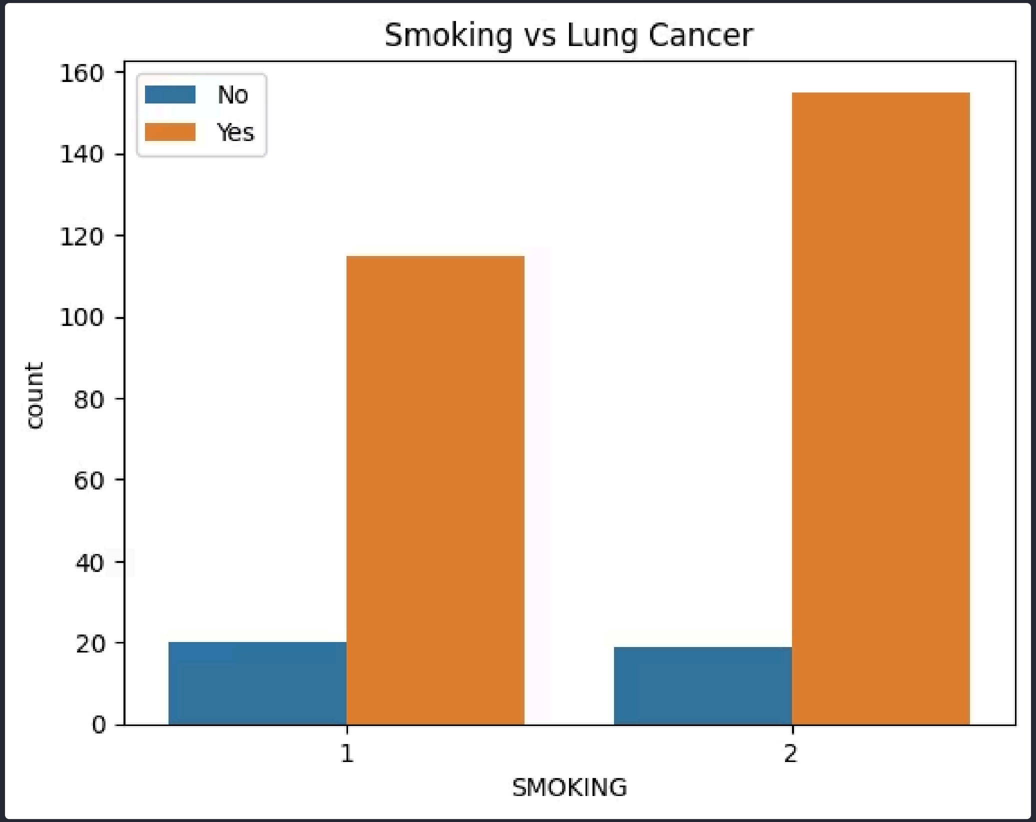
Sample Prediction Breakdown

The model outputs a risk score (0-1), indicating the likelihood of lung cancer. A threshold is then applied to classify it as positive or negative.

Sample Input: Gender=1, Age=60, and 13 symptom/health indicators (all value 2)
Output: Lung Cancer Prediction: YES

Key Feature Visualizations





Made with GAMMA

FUTURE

Future Scope and Next Steps

Our hackathon project is just the beginning. We envision a powerful, deployable tool with continuous improvements.

1

Model Refinement

Explore more advanced algorithms like Random Forest, Gradient Boosting, or even Neural Networks for enhanced predictive power.

2

Advanced Feature Engineering

Incorporate image data (CT scans, X-rays) using Convolutional Neural Networks (CNNs) to extract richer diagnostic features.

3

Real-World Deployment

Develop a user-friendly interface and integrate the model into clinical decision support systems for practical application.

4

Continuous Learning & Validation

Implement feedback loops for continuous model training with new data and rigorous validation with external datasets.