GROUP PROJECT

Medical Disease Prediction & Advice System

A Comprehensive ML-Based Health Advisory Platform

Course: SWE485 - Machine Learning | Domain: Healthcare

Presented By:
Waleed badkhin

Ahmad alawwad

Abdulaziz Alhelabi

Ahmad alharisi

Problem Statement

The Goal is to develop an intelligent medical advice system that predicts diseases from symptoms and provides actionable medical guidance.

- **Why This Problem? Addressing healthcare accessibility, enabling early detection, providing educational value, and serving as a robust, real-world ML application.**

Dataset Overview

- **Source: Kaggle - Disease Prediction Using Machine Learning**
- **Size: 4,920 patient cases, 132 binary symptom indicators**
- **Target: 41 unique diseases (e.g., Common Cold, Diabetes, Hepatitis, Heart Attack).**
- **Quality: High-quality, no missing values, well-structured, and ready for ML.**

Phase 1 - Data Exploration & Preprocessing

Key Activities & Findings

Dataset Analysis:

- Explored 4,920 rows and 133 columns.
- Identified 41 unique disease classes.
- Most common symptoms: Fatigue, cough, headache, fever patterns.

Preprocessing:

- Removed unnamed/empty columns.
- Verified binary feature format (0/1).
- Confirmed no missing values.
- Created a cleaned dataset for modeling.

Key Findings:

- High-quality, well-structured dataset.
- Binary features are ideal for classification.
- Relatively balanced class distribution.

Phase 2 - Supervised Learning

Algorithm Selection & Results

Selected Models:

1. Decision Tree: Interpretable, good for binary features.
2. Random Forest: Ensemble method, robust, known for high accuracy.
3. Naïve Bayes (BernoulliNB): Perfect fit for binary features, fast baseline.

Implementation:

- Training Set: 4,920 samples
- Testing Set: 42 samples
- Features: 132 binary symptoms
- Evaluation: Accuracy, Precision, Recall, F1-Score, Cross-Validation.

Phase 2 - Supervised Learning

Model performance

	Test Accuracy	Cross-Validation	Result
Decision Tree	88.1%	88.3%	Moderate performance
Random Forest	97.9%	100%	SELECTED - Best balance
Naïve Bayes	100%	100%	Perfect on clean dataset

Key Findings:

- Random Forest selected for the best balance of accuracy (97.6%) and robustness.
- Feature importance analysis revealed critical symptoms.
- Cross-validation confirmed model stability.

Phase 3 - Unsupervised Learning

Clustering Approach & Results
Algorithm: K-Means Clustering
Objective: Discover hidden patterns in symptom data (without disease labels).
Methodology:

- Removed the prognosis label, using 132 symptom features.
- Applied PCA (Principal Component Analysis) for dimensionality reduction.
- Selected K = 5 clusters using the Elbow Method.

Phase 3 - Unsupervised Learning

Results:

- 5 Clusters Identified: General, Respiratory, Dermatological, Digestive, Mixed.
- Evaluation Metrics:
  - Silhouette Score: 0.1942
  - BCubed Precision: 0.1210
  - BCubed Recall: 98.8% (high consistency)

Integration & Value:

- Groups patients by symptom patterns.
- Narrows down disease possibilities before supervised prediction.
- Enables personalized advice based on symptom clusters.
- Pipeline: Symptoms $\rightarrow$ Clustering $\rightarrow$ Supervised Model $\rightarrow$ Generative AI $\rightarrow$ Advice.

Phase 4 - Generative AI Integration

Objective & Implementation
Goal: Enhance the system with natural language generation for detailed medical explanations and actionable recommendations.
API Used: OpenAI GPT / Hugging Face (Mistral-7B-Instruct)
Approach: Tested 4 prompt templates (Simple, Structured, Conversational, Risk-Focused).

Phase 4 - Generative AI Integration

Criterion	Weight	Template 2 Score
Safety	30%	95%
Actionability	25%	90%
Comprehensiveness	20%	95%
Readability	15%	70%
Structure	10%	95%
Overall		91.5/100 ⭐

Key Findings:

- Random Forest selected for the best balance of accuracy (97.6%) and robustness.
- Feature importance analysis revealed critical symptoms.
- Cross-validation confirmed model stability.

Phase 4 - Generative AI Integration

Selected: Template 2 - Structured Medical Format
Why Selected?

- Highest safety score (95%) - includes medical disclaimers.
- Best comprehensiveness (95%) - covers all medical aspects.
- Excellent actionability (90%) - clear recommendations.
- Professional format - mirrors medical documentation.
- Easy integration with the supervised model.

System Integration & Overall Results

Complete System Pipeline

User Symptoms → Phase 3: Clustering (Patient Profile) and Phase 2: Supervised (Disease Prediction, 97.6% accuracy) → Phase 4: Generative AI (Detailed Advice) → Final Output: Disease + Recommendations + Actions

Integration Benefits:

- Layered analysis validates predictions.
- Comprehensive output (prediction + explanation).
- Cluster-based personalization.
- Professional medical consultation format.

Overall Results Summary

Phase	Component	Key Metric	Result
Phase 1	Data Quality	Missing Values	0%
Phase 2	Supervised Learning	Accuracy	**97.60%**
Phase 3	Clustering	BCubed Recall	98.80%
Phase 4	Generative AI	Template Score	91.50%

Key Achievements:

- 97.6% disease prediction accuracy (Random Forest).
- 5 meaningful symptom clusters identified.
- Professional structured medical advice format.
- Complete end-to-end pipeline.