Disease Prediction Using Machine Learning

-By
Prerna Dureja (229301453)
Avishi Semwal (229301154)

Introduction

- AI and ML have transformed healthcare.
- Disease prediction is vital for early diagnosis.
- Traditional diagnosis can be slow and error-prone.
- Aim: **Develop an intelligent ML system using Decision Tree Classifier.**

Dataset Description

- Training Dataset: 4920 rows, 133 columns (132 symptoms + 1 disease label)
- Testing Dataset: 420 rows, same structure
- Binary encoding of symptoms (1 = present, 0 = absent)
- Diseases are labeled under the **prognosis** column

Objective

- Build a machine learning model for accurate disease prediction.
- Use symptoms provided by the user as input.
- Improve prediction accuracy using **decision tree algorithm**

Problem Statement

- Given a list of symptoms, determine the most probable disease.
- Minimize false predictions by using effective preprocessing and tuning.
- Provide an interface where users can select symptoms easily.

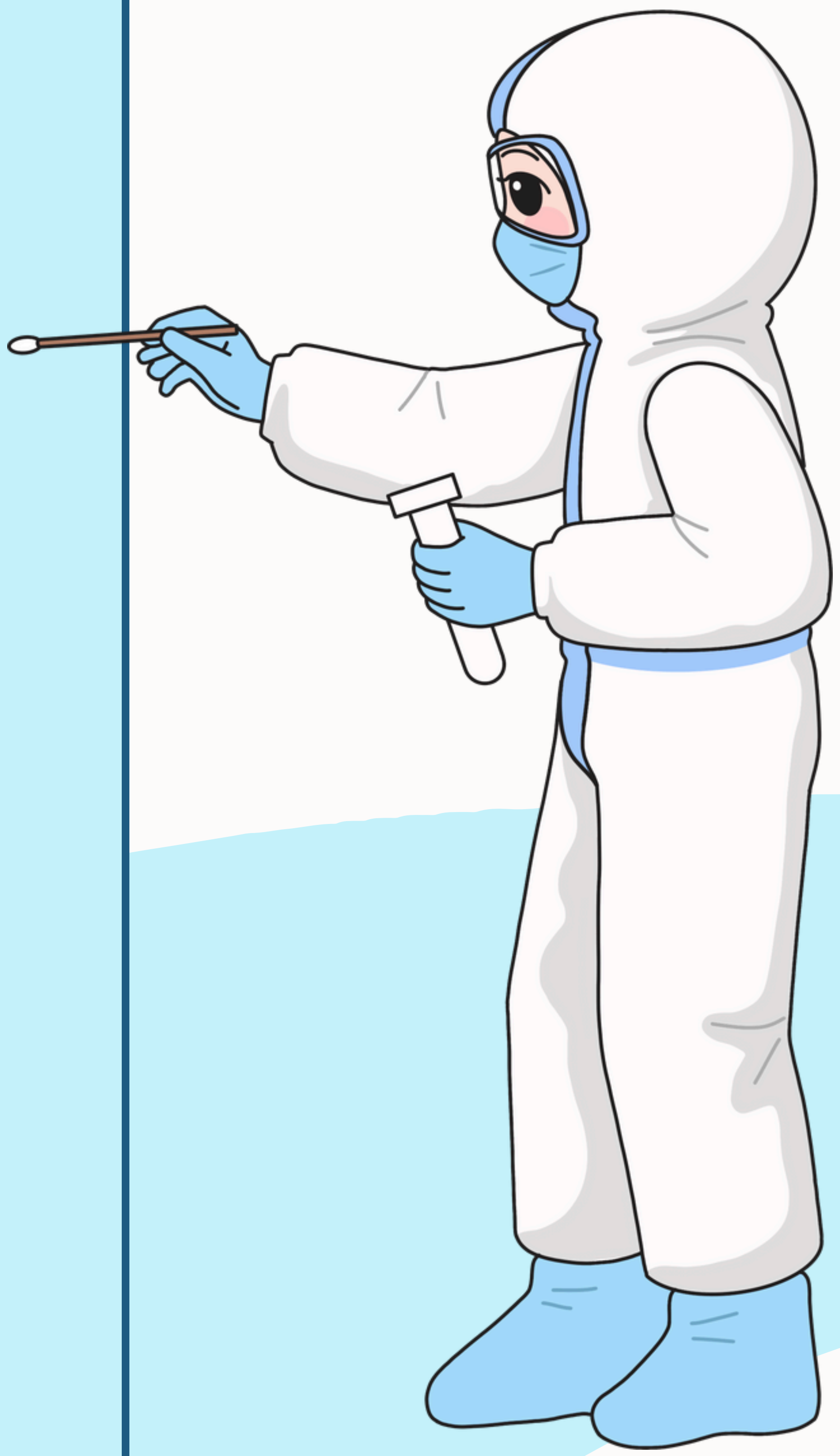
ML Algorithm - Decision Tree Classifier

- A supervised learning model based on tree-like structure.
- Splits data on features that result in the most information gain.
- Simple, interpretable, and suitable for small to medium-sized datasets.

Why Decision Tree?

- Our dataset uses binary symptom values (0 or 1), which Decision Trees naturally handle without the need for scaling or complex transformations.
- Provides a foundation to benchmark against advanced ensemble models like Random Forests or Gradient Boosting in the future.

Instruments Equipment Protection and security



Mathematical Foundation

- **Entropy:**

The impurity of a dataset is given by: $H(S) = -\sum pilog2pi$

- **Information Gain:**

The reduction in entropy when splitting on an attribute:
$IG(S,A) = H(S) - \sum(| Sv | / | S |)H(Sv)$

- **Gini Impurity:**

An alternative to entropy, calculated as:
$Gini = 1 - \sum pi^2$

Data Preprocessing

- Handled missing values using heuristics or default values.
- Encoded categorical values using label encoding.
- Split the dataset into training and testing sets.
- Balanced dataset ensured no class dominance.

Hyperparameter Tuning

- Criterion: 'gini' selected for impurity calculation
- Max Depth: Optimized to prevent overfitting
- Min Samples Split: Avoids creation of branches with very few samples
- Used Grid Search for optimal parameters

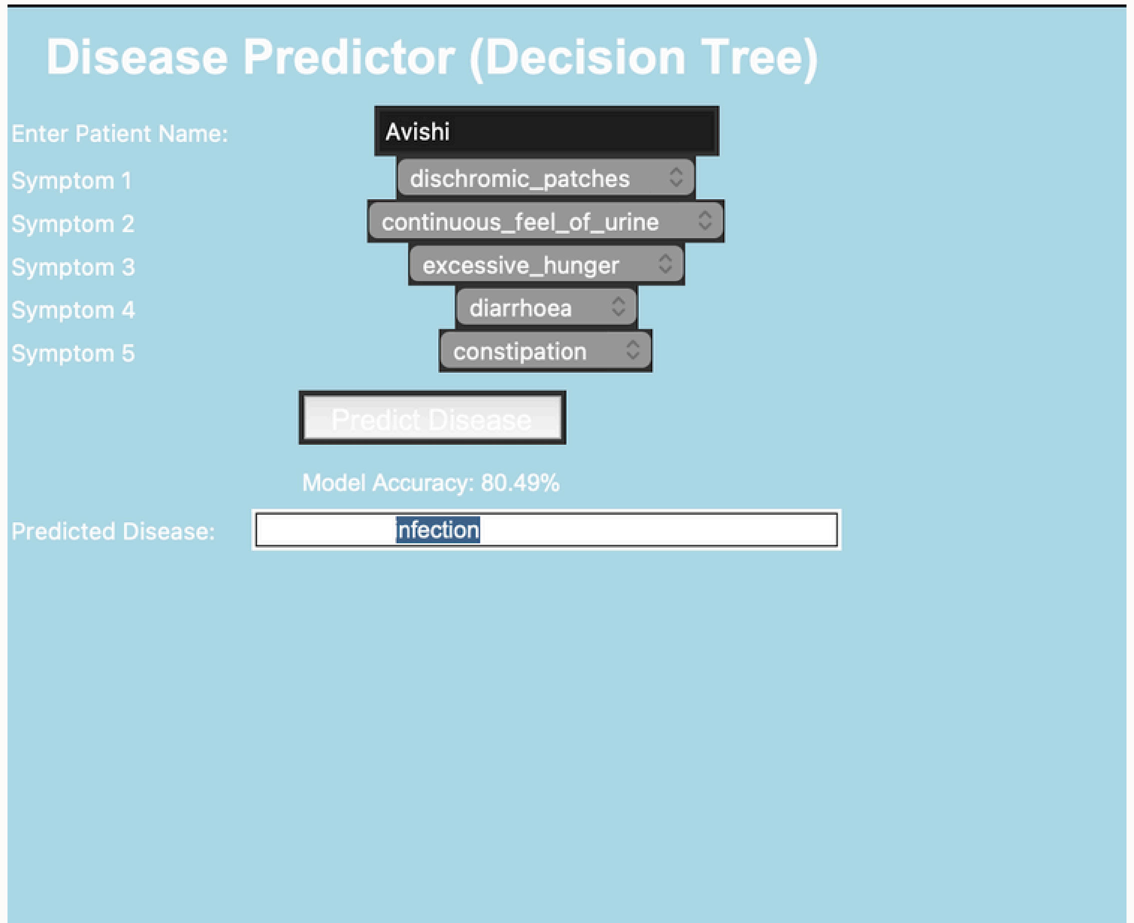
Pruning Techniques

-Pre-pruning:

- Restrict tree depth
- Minimum sample per leaf

-Post-pruning:

- Prune after full tree is built to reduce complexity
- Reduces overfitting and improves generalization



Disease Predictor (Decision Tree)

Enter Patient Name: Avishi

Symptom 1 dischromic_patches

Symptom 2 continuous_feel_of_urine

Symptom 3 excessive_hunger

Symptom 4 diarrhoea

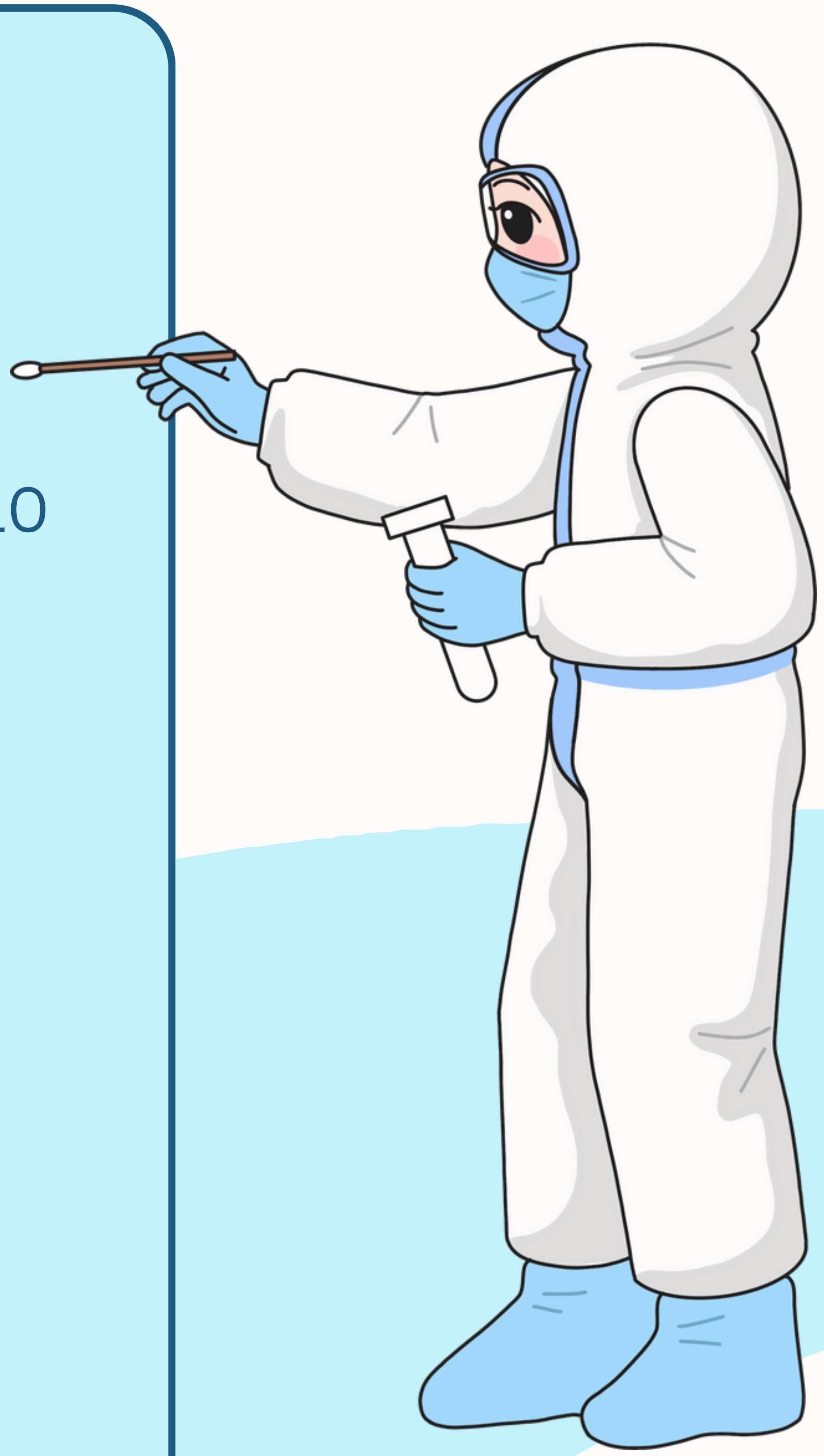
Symptom 5 constipation

Model Accuracy: 80.49%

Predicted Disease: nfection

Evaluation Metrics

- Confusion Matrix: TP: 90, TN: 50, FP: 29, FN: 10
- Accuracy: 80.49%
- Precision: 75.63%
- Recall: 90%
- F1 Score: 82.2%

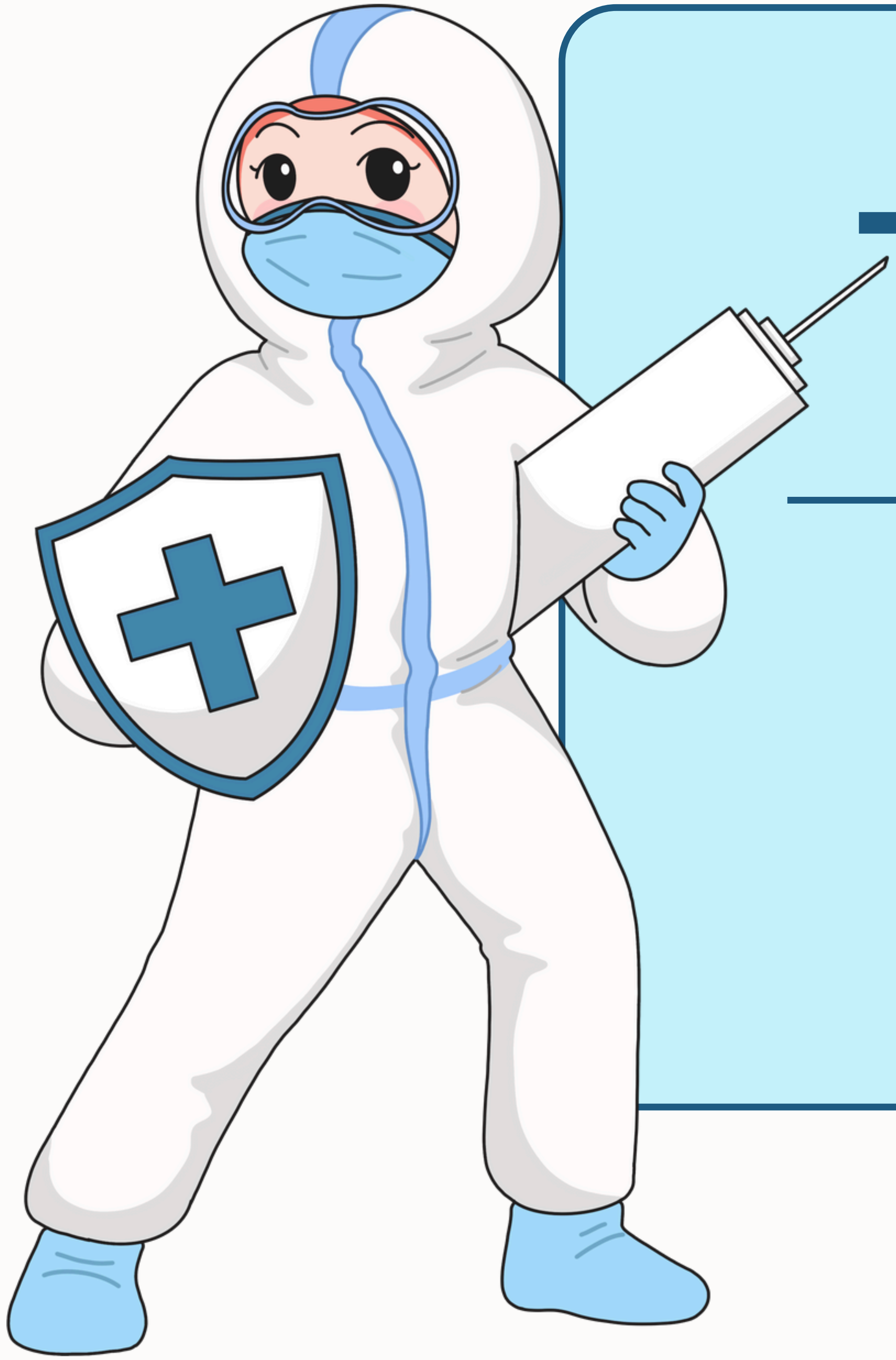


Limitations and Future Work

- Limited to dataset diseases only
- Accuracy can be improved using ensemble methods
- Future enhancements:
  - Add more symptoms
  - Use Random Forest / XGBoost
  - Web-based GUI

Conclusion

- Successfully built a disease prediction system using Decision Trees
- Achieved decent performance metrics
- Potential for real-world application with improvements



THANK YOU!