International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 09 | Sept - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

# DISEASES FORECASTING USING DECISION TREE

Dr Shivandappa, Dr Narendra Kumar S, Bazilla Wani, Bhumika Mandolkar

R.V College of Engineering, Bengaluru-560059

## ABSTRACT

This paper presents a unambiguous approach to health forecasting using a decision tree classifier, which is implemented in Python. We developed a model to categorize illness according to symptoms and provide corresponding preventive measures. The decision tree model was trained on a dataset where symptoms were represented as categorical variables rather than binary, making the dataset more intuitive and flexible. The model processes user inputs, converting them into numerical format, and predicts the most likely disease. Additionally, we visualize the decision tree to illustrate the decision-making process from symptoms to disease prediction, facilitating a better understanding the behaviour of model. This approach not only demonstrates the effectiveness of decision trees in medical diagnostics but also provides actionable precautions tailored to the predicted disease, enhancing both diagnosis and prevention strategies.

**Keywords:** Decision Tree Classifier, Health forecasting, Numerical format, Symptom Analysis, Data Encoding, Visualization, Machine Learning, Python Programming, Medical diagnostics

## INTRODUCTION

Healthcare machine-learning models have proven to be a boon to the industry, swiftly assisting physicians and medical staff in managing an ever-growing patient population. While these models cannot completely diagnose chronic or minor illnesses without professional intervention, they can make situations easier for both healthcare professionals and patients seeking help investigation based on symptoms. There are significant reasons to show how these models are proving effective in the industry. Individuals can seek early help by analysing their symptoms with the help of these models and getting the necessary assistance in the early stages of any illness they may have. Disease prediction models can enable personalized medicine by taking into account an individual's specific characteristics, such as genetic information, lifestyle factors, and medical history. These models can empower individuals by providing personalized risk assessments and insights into their health status, motivating them to adopt healthier lifestyles, adhere to preventive measures, and actively engage in their healthcare management, leading to improved health outcomes. It's important to remember that disease prediction models should be developed and validated using high-quality data while considering ethical and privacy considerations. They should also be continually refined and updated as new data and knowledge become available to ensure their accuracy, reliability, and relevance in real-world healthcare settings.

## METHODOLOGY

The methodology section outlines the approach and steps taken to develop and evaluate the disease prediction model using a decision tree classifier. This process includes dataset preparation, model training, prediction, and visualization. Here's a detailed breakdown:

### 1. Data Collection and Preparation

The initial step involved constructing a synthetic dataset to simulate the relationship between various symptoms and diseases. The dataset included the following elements:

- **Symptoms:** A set of common symptoms such as fever, cough, fatigue, headache, and sore throat.
- **Diseases:** A small set of diseases—Flu, Cold, and Migraine—each associated with a specific combination of symptoms.
- **Data pruning of decision tree:** is a technique used to simplify decision trees by removing unnecessary parts, preventing overfitting and improving model performance on new data.

The dataset was structured such that each disease was associated with a unique combination of symptoms. For instance, the Flu was characterized by symptoms like fever, cough, and fatigue, while a Cold was associated with fever, cough, and sore throat. The dataset was then organized into a feature matrix (X) representing the symptoms and a target vector (Y) representing the corresponding disease labels.

### 2. Data Encoding

To facilitate the use of symptoms as features in a machine-learning model, we converted symptom names into a binary matrix. This step involved:

- **Symptom Extraction:** Identifying all unique symptoms from the dataset.
- **Binary Encoding:** Each symptom was represented as a binary value—1 if the symptom was present and 0 if it was absent in a particular instance.

This encoding allowed the model to process symptoms as numerical data, enabling it to distinguish between the presence and absence of each symptom for each disease.

### 3. Model Selection and Training

*A* Decision Tree Classifier was selected for this task due to its interpretability and suitability for small datasets. The decision tree algorithm operates by:

- **Recursive Binary Splitting:** The decision tree repeatedly splits the data into subsets based on the most significant symptom (feature) at each node. The significance of a symptom is measured by its ability to increase the homogeneity of the resulting subsets.
- **Node Structure:** Each internal node of the tree represents a decision based on the value of a particular symptom. The node's branches correspond to the possible outcomes of that decision (e.g., the presence or absence of the symptom).
- **Leaf Nodes:** The terminal nodes of the tree, known as leaf nodes, represent the predicted disease. Each leaf node is associated with a specific disease label, indicating the final classification outcome for a given set of symptoms.

The model was trained using the DecisionTreeClassifier from the sci-kit-learn library, which automatically determined the optimal splits and constructed the tree based on the training data.

**4. User Input and Prediction**

After training, the model was designed to take user input in the form of answers to symptom-related questions (e.g., "Do you have a fever?"). The user's responses were encoded into a binary format, matching the format used during training. This encoded input was then passed through the trained decision tree to generate a prediction:

- **Traversal of Nodes:** The input vector traverses the tree from the root node, following the branches corresponding to the user's responses. The traversal continues until it reaches a leaf node, which provides the predicted disease.

**5. Visualization of the Decision Tree**

To enhance the interpretability of the model, the decision tree was visualized using the plot_tree function from scikit-learn. The visualization provided a clear representation of:

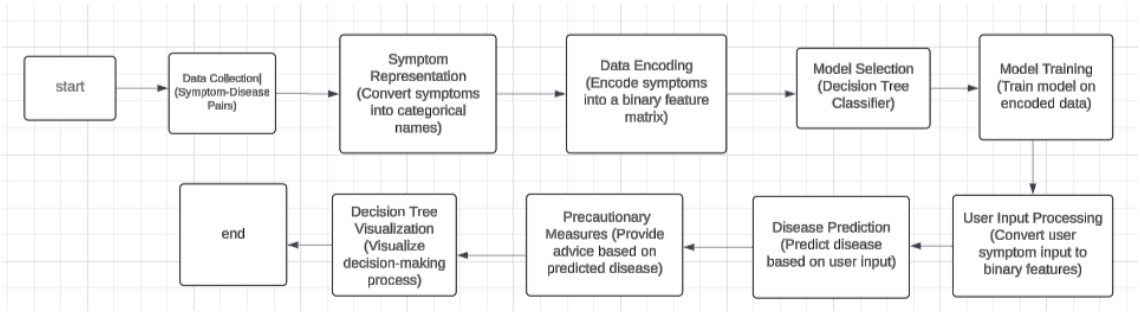
- **Feature Splits:** Each node in the tree displayed the symptom that was used to split the data.
- **Class Labels:** The leaf nodes displayed the predicted diseases, illustrating the path from symptoms to diagnosis.
- **Node Information:** Each node included information about the number of samples that reached that node and the distribution of diseases among those samples.

This visualization served as an explanatory tool, allowing users and reviewers to understand the decision-making process behind the model's predictions.

**6. Evaluation and Precautionary Recommendations**

After obtaining a prediction, the model provided a set of precautionary measures tailored to the predicted disease. These recommendations were predefined based on standard medical advice for each disease in the dataset. The system's ability to offer both predictions and actionable advice highlights its practical applicability in real-world scenarios, particularly in preliminary disease screening and prevention.

Flowchart: The proposed design of this disease prediction model can be depicted by the below flowchart




The future scope for disease predictor models is vast and holds significant potential for advancements in healthcare. Here are some potential areas of development and improvement for disease predictor models: Integration of Big Data and Advanced Technologies: Multi-Disease Prediction, Real-Time Monitoring and Early Warning Systems, Explain

ability and interpretability, Ethical Considerations, and Privacy Protection. Overall, the future scope for disease predictor models lies in advancing data integration, incorporating advanced technologies, personalizing healthcare interventions, and addressing global health challenges. By continuously improving the accuracy, interpretability, and ethical considerations of these models, they can play a pivotal role in early detection, preventive healthcare, and improving patient outcomes.

## CONCLUSION

In conclusion, a disease prediction system using a decision tree algorithm can be an effective approach for predicting and classifying diseases based on input features. Decision trees are a popular machine learning algorithm that builds a tree-like model to make decisions by recursively partitioning the data based on feature values. By utilizing a decision tree in a disease prediction system, advantages like interpreting ability, efficiency, and handling of categorical data can be achieved. A disease prediction system offers several advantages that can greatly benefit healthcare professionals, patients, and the overall healthcare system. It can greatly benefit individuals who are unaware and uncertain about the symptoms they are experiencing. Early diagnosis can help them find appropriate medical help that fits their needs. Medical staff can also use this model to book appropriate appointments based on this analysis, leading to improved and efficient treatment planning. Additionally, it can contribute to healthcare awareness. It's important to note that while disease prediction systems offer numerous advantages, they should always be used in conjunction with clinical judgment and medical expertise.

## LITERATURE SURVEY

| Year | Objective | Link |
|---|---|---|
| 2023 | DISEASE PREDICTION SYSTEM USING DECISION TREE CLASSIFIER | https://www.irjmets.com/uploadedfiles/paper/issue_6_june_2023/42422/final/fin_irjmets1687405587.pdf |
| 2008 | Decision Tree Algorithms Predict the Diagnosis and Outcome of Dengue Fever in the Early Phase of Illness | https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2263124/#abstract-a.k.b.xtitle |
| 2022 | Prediction of diagnosis and prognosis of COVID-19 disease by blood gas parameters using decision trees machine learning model: a retrospective observational study | https://pubmed.ncbi.nlm.nih.gov/34677154/ |
| 2024 | Machine learning-assisted prediction of pneumonia based on non-invasive measures | https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9371749/ |
| 2014 | Migraine diagnosis by using artificial neural networks and decision tree techniques | https://www.researchgate.net/figure/Migraine-and-probable-migraine-decision-tree-obtained-by-using-Gini-algorithm_fig1_276137207 |
| 2022 | Disease Prediction using Machine Learning | https://ijcrt.org/papers/IJCRT2203398.pdf |

| | | |
|---|---|---|
| 2021 | Efficient Prediction of Infectious and Non-Infectious Diseases using Decision Tree Classifier Algorithm | https://www.riverpublishers.com/pdf/ebook/chapter/RP_9788770040723C154.pdf |
| 2020 | Forecasting Weekly Influenza Outpatient Visits Using a Two-Dimensional Hierarchical Decision Tree Scheme | https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7369891/ |
| 2020 | Prediction of adverse drug reactions using decision tree modelling | https://pubmed.ncbi.nlm.nih.gov/20220749/ |
| 2022 | Disease-Prediction-Decision-Tree | https://www.kaggle.com/code/danushkumarv/disease-prediction-decision-tree |