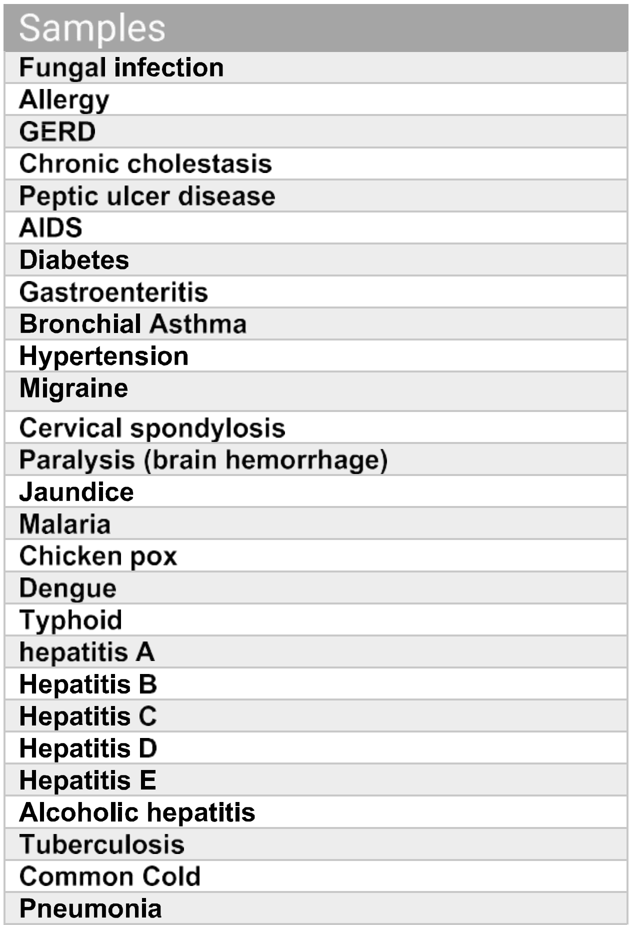
Disease Diagnosis

Problem Statement : Build and evaluate the machine learning model for disease diagnosis from patients' symptoms.

About Dataset:

The data set have be collected from various sources such as a file, database, and many other such sources. We have also used some free data sets which are present on the internet. Kaggle and UCI Machine learning Repository.



Samples
Fungal infection
Allergy
GERD
Chronic cholestasis
Peptic ulcer disease
AIDS
Diabetes
Gastroenteritis
Bronchial Asthma
Hypertension
Migraine
Cervical spondylosis
Paralysis (brain hemorrhage)
Jaundice
Malaria
Chicken pox
Dengue
Typhoid
hepatitis A
Hepatitis B
Hepatitis C
Hepatitis D
Hepatitis E
Alcoholic hepatitis
Tuberculosis
Common Cold
Pneumonia

Dimorphic hemmorhoids(piles)
Heart attack
Varicose veins
Hypothyroidism
Hyperthyroidism
Hypoglycemia
Arthritis
(vertigo) Paroymsal Positional Vertigo
Acne
Urinary tract infection
Psoriasis
Impetigo

No of features:132

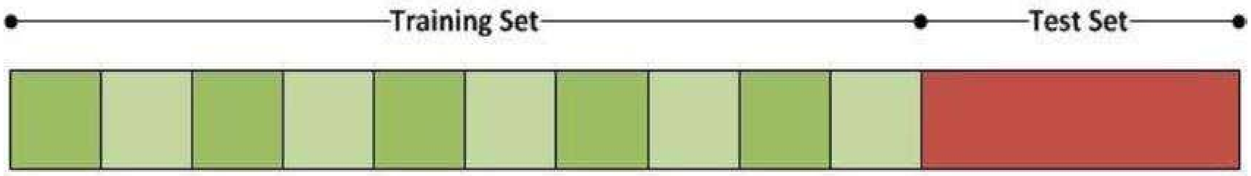
We can define the machine learning workflow stages as:

1. Gathering data
   collected data cannot be used directly for performing the analysis process as there might be a lot of missing data, extremely large values, unorganized text data or noisy data. Therefore, to solve this problem Data Preparation is done.

2. Data pre-processing
   Data pre-processing is one of the most important steps in machine learning. It is the most important step that helps in building machine learning models more accurately. Data pre-processing is a process of cleaning the raw data i.e. the data is collected in the real world and is converted to a clean data set.
   - Converting samples into numerical data as classifier model only works on numerical data
   - Missing data can be replaced by adding samples or deleting rows and columns.
3. Researching the model that will be best for the type of data
   Our main goal is to train the best performing model possible, using the pre-processed data.
   We used classification algorithms.

   - **Decision Trees**

4. Training and testing the model
   For training a model we initially split the model into 2 sections which are '**Training data**' and '**Testing data**'. During training the classifier only the training set is available. The test data set must not be used during training the classifier. The test set will only be available during testing the classifier.



Algorithm of the code:

1. Import Numpy, Pandas, Sci-kit Learn, Tkinter libraries.

2. Gathering the data.

3.Data visualization with help of Count graph, Accuracy graph, Decision tree with Gini index.

3. Data cleaning and processing is done. Data pre-processing can be done

using –

1. Conversion of data: As we know that Machine Learning models can only

handle

numeric features, hence categorical and ordinal data must be somehow

converted into

numeric features.

2. Ignoring the missing values: Whenever we encounter missing data in the

data set

then we can remove the row or column of data depending on our need.

3. Machine learning: If we have some missing data then we can predict what

data shall

be present at the empty position by using the existing data

4. Training and testing the model on data.

5. Decision tree is implemented.

6. Confusion matrix is implemented.

7. Accuracy of the model is found using confusion matrix by the above formula-

Accuracy (all correct / all) = TP + TN / TP + TN + FP + FN

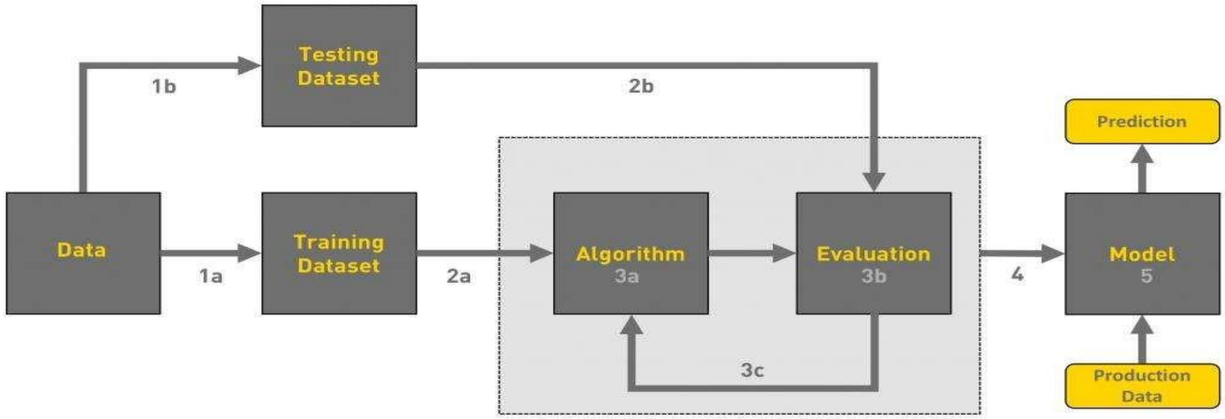
8. Model is evaluated.

Evaluation

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future.

Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right.

Data visualization can be done by using plotting graph. By taking on X-axis there will be prognosis .



Justification:

Why we use accuracy?

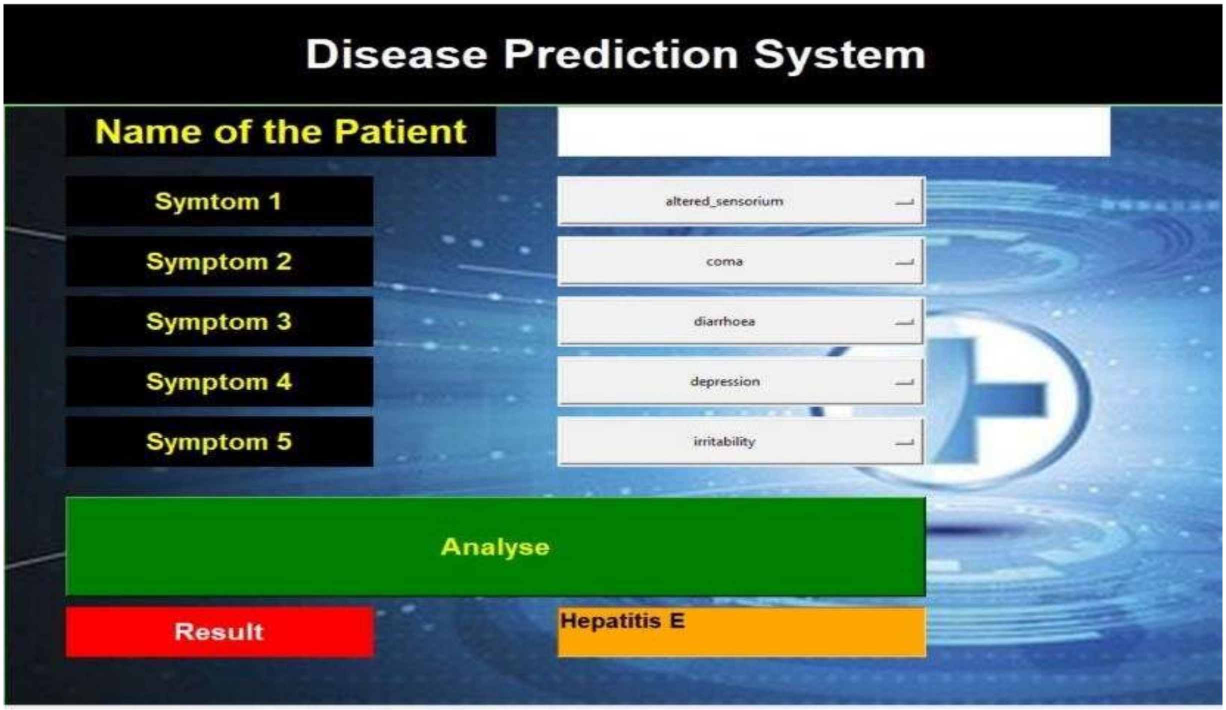
Companies use machine learning models to make practical business decisions, and **more accurate model outcomes result in better decisions.** The cost of errors can be huge, but optimizing model accuracy decresing that cost.

Good accuracy in machine learning is subjective. Accuracy is **to be ensuring that the information is correct and without any mistake**. accuracy is important because may the life of people depend on it .

Result:

Accuracy	94.53%

Here for example when we give some symptoms of diabetes as input in model , the model predicts the disease diabetes with 94.53% accuracy



Name of the Patient

Symtom 1: altered_sensorium

Symptom 2: coma

Symptom 3: diarrhoea

Symptom 4: depression

Symptom 5: irritability

Result: Hepatitis E

Conclusion:

We have Successfully implemented decision model using decision tree algorithm.

Thus, these ML model helps the doctors to predict the diseases with more accuracy and reduces the cost of testing.

This project aims to predict the disease on the basis of the symptoms.

The use of this ML model enabled the early detection of many maladies such as diabetes ,common cold ,tuberculosis ,pneumonia , heart attack ,etc.

As a system is based on software ,the user can use this system from anywhere and at any time.

In this model decision tree ,confusion matrix and more evaluation parameters are used to predict the disease.