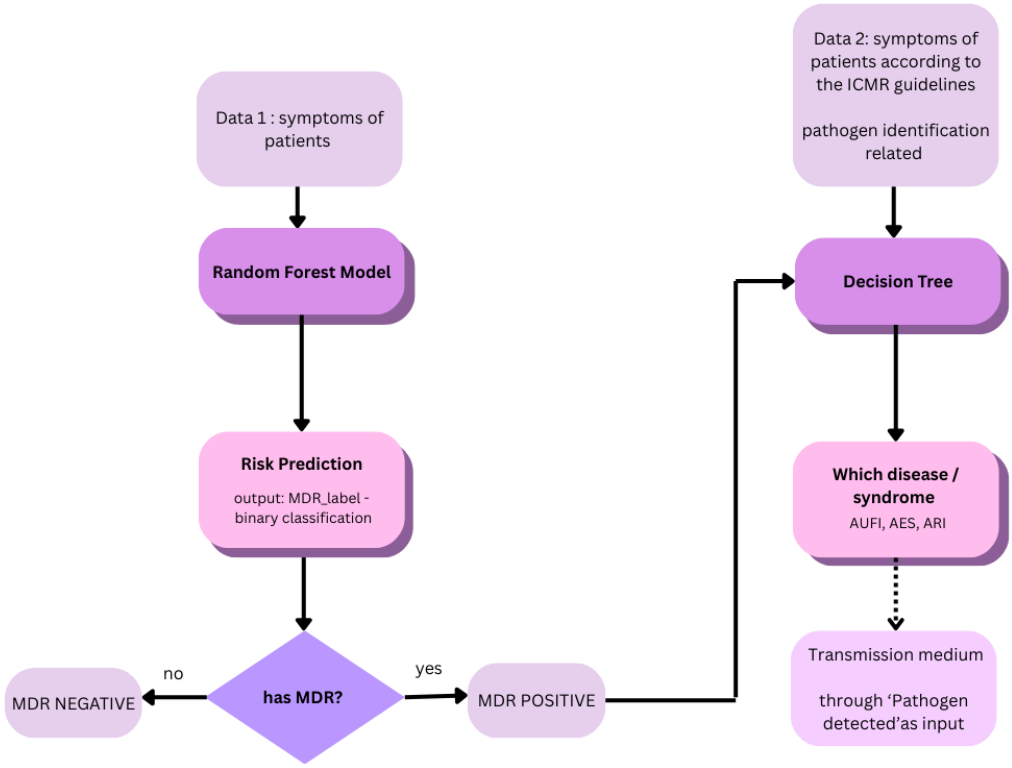
ML MODEL DOCUMENTATION



1. Random Forest Model

Data1:
This dataset contains clinical symptoms of the patient entered through the web/ application this input is used to predict the MDR risk.
Random Forest Classifier predicts the Risk and if the patient tests positive for MDR_label, then the type of MDR is decided using Decision Tree Model.

2. Decision Tree Classifier

Data2: Based on ICMR (Indian Council of Medical Research) Guidelines, we have another dataset that has features like 'cough', 'mental_wellbeing', 'fever_days', 'stool', etc – different from the data1.

This classifier classifies the MDR positive case into : 4 classes

1. Acute Undifferentiated Febrile Illness (AUFI)

Syndromic Surveillance of Infectious Diseases
Priority Pathogens To Be Tested For In Indian Clinical Settings

1. PRIORITY PATHOGENS CAUSING ACUTE FEBRILE ILLNESS

i. Acute Undifferentiated Febrile Illness (AUFI):

Case definition: Patient with fever ≥ 38.0°C or 100.4°F of ≤ 14 days duration without localization (i.e., without evidence of localized infection by history, physical examination, investigations such as complete blood count, biochemistry profile, urine analysis, or chest radiography) at the time of initial presentation.

2. Priority pathogens causing acute encephalitis syndrome(AES)

Syndromic Surveillance of Infectious Diseases
Priority Pathogens To Be Tested For In Indian Clinical Settings

2. PRIORITY PATHOGENS CAUSING ACUTE ENCEPHALITIS SYNDROME (AES)

Case Definition

A case of AES is defined by the National Center for Vector Borne Diseases Control (NCVBDC) as a person of any age, presenting with the below symptoms at any time of the year:

Acute onset of fever (≤7 days)
AND
Change in mental status (including symptoms such as confusion, disorientation, coma or inability to talk)
AND /OR
New onset of seizures (excluding simple febrile seizures)

3. Priority pathogens causing acute respiratory illness(ARI)

Syndromic Surveillance of Infectious Diseases
Priority Pathogens To Be Tested For In Indian Clinical Settings

3. PRIORITY PATHOGENS CAUSING ACUTE RESPIRATORY ILLNESS (ARI)

Case Definition

Influenza Like Illness (ILI) and Severe Acute Respiratory Infection (SARI) case definitions may be employed for surveillance studies for detection of influenza, SARS-CoV-2 and other viral aetiologies.

ILI is defined by the World Health Organization (WHO) as an acute respiratory infection with:

- measured fever of ≥ 38 C°
- and cough;
- with onset within the last 10 days

SARI is defined by WHO as an acute respiratory infection with:

- history of fever or measured fever of ≥ 38 C°
- and cough;
- with onset within the last 10 days,
- and requires hospitalization.

4. Other

This then, can further predict transmission medium.

COMPARISION STUDY

Logistic Regression

```
y_pred = clf.predict(X_test)
print(classification_report(y_test, y_pred))
```

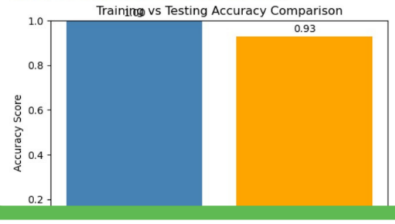
	precision	recall	f1-score	support
0	0.95	0.99	0.97	278
1	0.75	0.27	0.40	22
accuracy			0.94	300
macro avg	0.85	0.63	0.68	300
weighted avg	0.93	0.94	0.93	300

XGBoost Performance

```
Classification report (Test):
              precision    recall  f1-score   support

           0       0.97      1.00      0.99       378
           1       0.92      0.55      0.69        22

    accuracy                           0.97       400
   macro avg       0.95      0.77      0.84       400
weighted avg       0.97      0.97      0.97       400
```

Random Forest Performance

```
Training Accuracy : 1.0
Testing Accuracy  : 0.93

Model Generalization Looks Good: No major overfitting.
```



On checking with multiple models for data1,

We concluded that Random forest has the best performance because of the following reasons:

1. It is resistant to overfitting, because it uses bagging technique.
2. It handles large amount of data well.
3. It is effiecient for non-linear data.
4. It is not as complex as XGBoost.

And since the rules for the data2 classification were clearly available from the ICMR guildelines, we used Decision Trees as it would be helpful for Feature seection and Threshold value.