Artificial Intelligence Project
DSCI-6612

Team members
Namitha Nagaraju
Syeda Fatiha Buttul

Under the Guidance of
Vahid Behzadan

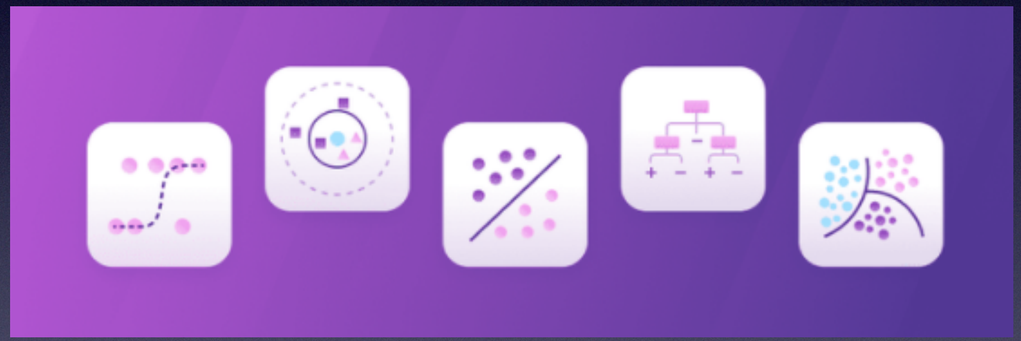
Date
10 Decemeber 2021

PROJECT TITLE

Disease prediction using Machine Learning

MOTIVATION

- Machine learning methods can be used to create models obtained by training them on a known dataset to predict new information.
- Medicine is a very critical field of science which needs high attention to details while modelling, to predict this new information.
- With the advancement of technology, leveraging the advantages of the machine's high precision, speed and accuracy can help save a lot of time, effort and money.
- This project aims at providing initial diagnosis of deceases when a patient enters his symptoms into the systems. This initial information regarding a probable medical problem can help patients save their money by not consulting a doctor who usually take a lot of money during first consulting sessions. They can also research and then go to a professional for further treatment.



Dataset

https://www.kaggle.com/kaushil268/disease-prediction-using-machine-learning

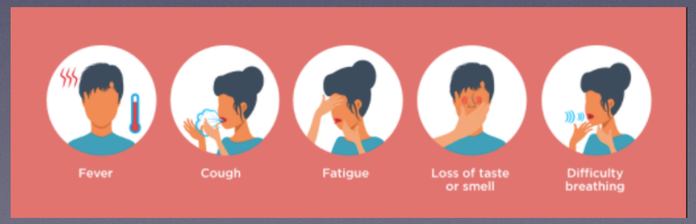
- 132 COLUMNS ARE SYMPTOMS USED TO CLASSIFY 41 KNOWN DISEASES.
- THE TEST DATASET HAS 13 KB OF DATA AND THE TRAIN DATA HAS 1.3 MB OF DATA.

itching	skin_rash	nodal_skin_e	continuous_s	shivering	chills	joint_pain	stomach_pai	acid
1	1	1	0	0	0	0	0	
0	0	0	1	1	1	0	0	
0	0	0	0	0	0	0	1	
1	0	0	0	0	0	0	0	
1	1	0	0	0	0	0	1	
0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	
0	0	0	0	0	1	0	0	
1	1	0	0	0	0	0	0	
0	1	0	0	0	1	1	0	
0	0	0	0	0	1	0	0	

Snapshot of dataset indicating the different symptoms

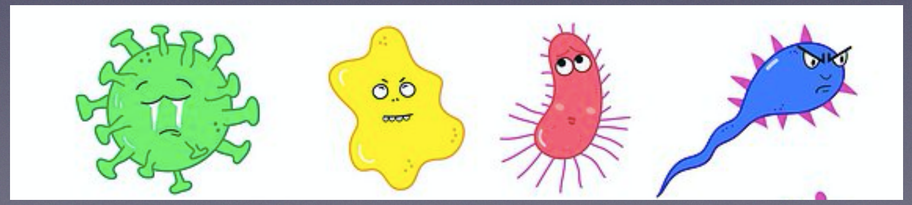
DEPENDENT VARIABLES (GIVEN)

- **itching**
- **skin rash**
- **shivering**
- **chills**
- **anxiety**
- **lethargy**
- **blister**
- **nausea**
- **caugh**
- **weight gain**
- **cold hands**
- **mood swing**
- **ETC**



INDEPENDENT VARIABLES (TO PREDICT)

- **Allergy**
- **GERD**
- **Chronic cholestasis**
- **Drug Reaction**
- **Peptic ulcer diseae**
- **AIDS**
- **Diabetes**
- **Gastroenteritis**
- **Bronchial Asthma**
- **Hypertension**
- **Migraine**
- **ETC**



APPROACH

ALL THE BELOW CLASSIFICATION ALGORITHMS WILL BE USED TO TRAIN THE DATA. THE BEST OF THESE MODELS WILL BE USED TOIMPLEMENT AN END TO END APPLICATION HELPFUL FOR A PATIENT TO INITIALLY DIAGNOSE THIER PROBLEM BEFORE GOING TO A SPECIALISED DOCTOR.

Logistic Regression
Random Forest
Support Vector Machine
KNeighbors
Decision Tree
Ada Boost
Bagging
Gradient Boosting

DELIVERABLES

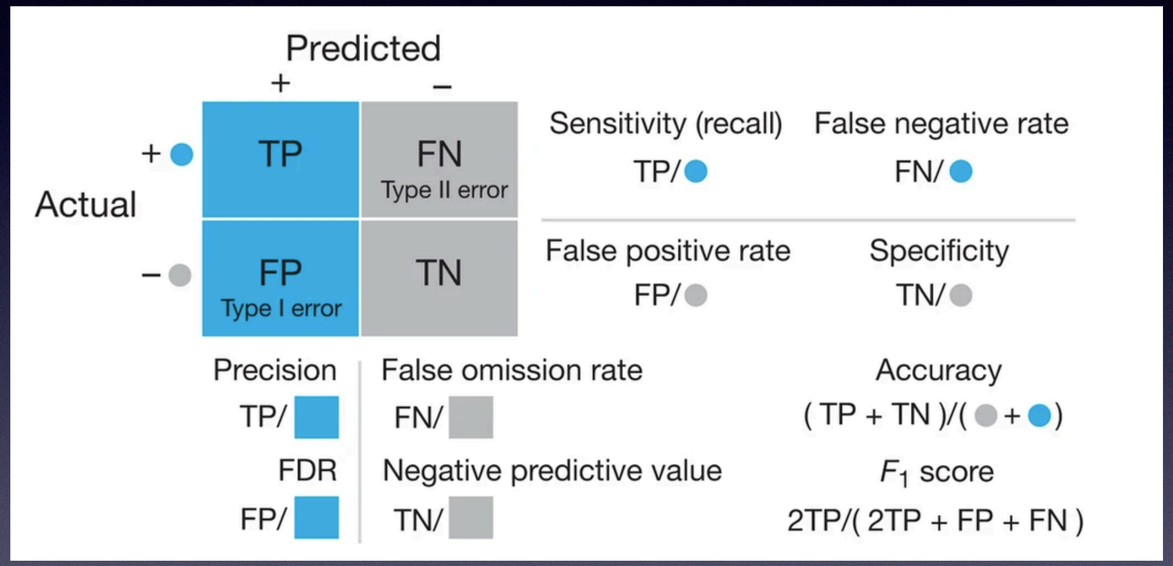
JUPITER NOTEBOOK: Presentation outlining

- Data Cleaning
- Data Exploration
- Implementation of the machine learning algorithms
- Results of different methods
- Comparison of different methods
- Evaluation methods
- Selection of best method

•

Evaluation method

CONFUSION MATRIX



Blue and gray circles indicate cases known to be positive (TP + FN) and negative (FP + TN), respectively, and blue and gray backgrounds/squares depict cases predicted as positive (TP + FP) and negative (FN + TN), respectively. Equations for calculating each metric are encoded graphically in terms of the quantities in the confusion matrix. FDR, false discovery rate

Source: https://www.nature.com/articles/nmeth.3945/figures/1

Results

TRAIN CV ACCURANCY: 1.000

TEST ACCURACY: 1.000

CONFUSION MATRIX:

```
[[50  0  0 ...  0  0  0]
 [ 0 50  0 ...  0  0  0]
 [ 0  0 50 ...  0  0  0]
           ...
 [ 0  0  0 ... 50  0  0]
 [ 0  0  0 ...  0 50  0]
 [ 0  0  0 ...  0  0 50]]
```

CLASSIFICATION REPORT

	precision	recall	f1-score	support
0	1.00	1.00	1.00	50
1	1.00	1.00	1.00	50
2	1.00	1.00	1.00	50
3	1.00	1.00	1.00	50
4	1.00	1.00	1.00	50
5	1.00	1.00	1.00	50
6	1.00	1.00	1.00	50
7	1.00	1.00	1.00	50
8	1.00	1.00	1.00	50
9	1.00	1.00	1.00	50
10	1.00	1.00	1.00	50
11	1.00	1.00	1.00	50
12	1.00	1.00	1.00	50
13	1.00	1.00	1.00	50
14	1.00	1.00	1.00	50
15	1.00	1.00	1.00	50
16	1.00	1.00	1.00	50
17	1.00	1.00	1.00	50
18	1.00	1.00	1.00	50
19	1.00	1.00	1.00	50
20	1.00	1.00	1.00	50
21	1.00	1.00	1.00	50
22	1.00	1.00	1.00	50
23	1.00	1.00	1.00	50
24	1.00	1.00	1.00	50
25	1.00	1.00	1.00	50
26	1.00	1.00	1.00	50
27	1.00	1.00	1.00	50
28	1.00	1.00	1.00	50
29	1.00	1.00	1.00	50
30	1.00	1.00	1.00	50
31	1.00	1.00	1.00	50
32	1.00	1.00	1.00	50
33	1.00	1.00	1.00	50
34	1.00	1.00	1.00	50
35	1.00	1.00	1.00	50
36	1.00	1.00	1.00	50
37	1.00	1.00	1.00	50
38	1.00	1.00	1.00	50
39	1.00	1.00	1.00	50
40	1.00	1.00	1.00	50

accuracy			1.00	2050
macro avg	1.00	1.00	1.00	2050
weighted avg	1.00	1.00	1.00	2050

Thank you