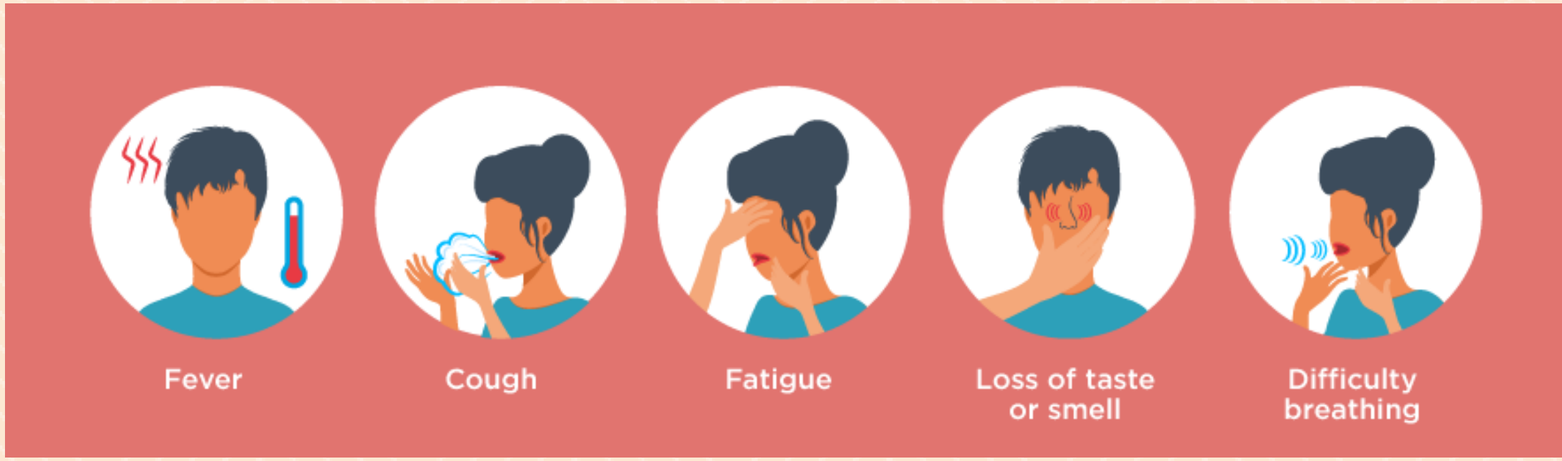
DISEASE SYMPTOM PREDICTION

By Ashwin Shibu and Siddharth Choudhury

Introduction

Various diseases affect people today because of the environment and their lifestyle choices. Predicting sickness early on so becomes a crucial task. However, based just on symptoms, doctors find it difficult to make exact predictions. The hardest challenge is to accurately predict diseases.

The development of a Disease Symptoms Prediction Application based on machine learning (ML) algorithms for illness prediction can aid in a more accurate diagnosis than the current methods. Using Supervised machine-learning techniques, we created a disease prediction system



The correct prediction of disease is the most challenging task. The system processes the symptoms provided by the user as input and gives the output as the probability of the disease. Naïve Bayes classifier is used in the prediction of the disease which is a supervised machine learning algorithm. The probability of the disease is calculated by the Naïve Bayes algorithm.

With an increase in biomedical and healthcare data, accurate analysis of medical data benefits early disease detection and patient care. After general disease prediction, this system able to gives the risk associated with general disease which is lower risk of general disease or higher. By using linear regression and decision tree we are predicting diseases like Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis.

Calculate Performance Evaluation

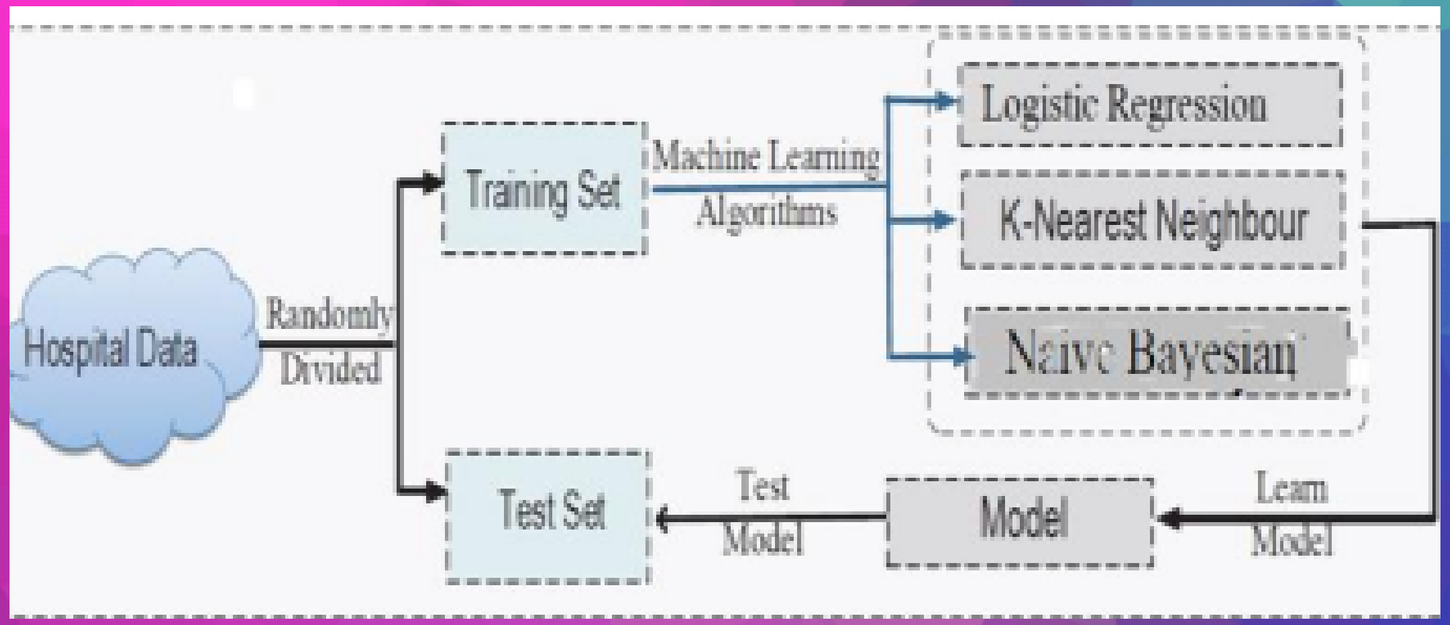
accuracy -:

$$\frac{TruePositive+TrueNegative}{TruePositive+TrueNegative+FalsePositive+FalseNegative}$$

$$\text{Precision} = \frac{TruePositive}{TruePositive+FalsePositive}$$

$$\text{Recall} = \frac{TruePositive}{TruePositive+FalseNegative}$$

$$\text{F1-Measure} = \frac{2\times precision\times recall}{precision+recall}$$



Research Papers

- Pingale, Kedar, Sushant Surwase, Vaibhav Kulkarni, Saurabh Sarage, and Abhijeet Karve. "Disease prediction using machine learning." International Research Journal of Engineering and Technology (IRJET) 6 (2019): 831-833.
- Arora, P., Boyne, D., Slater, J. J., Gupta, A., Brenner, D. R., & Druzdzel, M. J. (2019). Bayesian networks for risk prediction using real-world data: a tool for precision medicine. Value in Health, 22(4), 439-445.
- Learning, M. (2017). Heart disease diagnosis and prediction using machine learning and data mining techniques: a review. Advances in Computational Sciences and Technology, 10(7), 2137-2159.

Cleaning the Dataset

- Step 1: Remove duplicate or irrelevant observations. Remove unwanted observations from your dataset, including duplicate observations or irrelevant observations. ...
- Step 2: Fix structural errors. ...
- Step 3: Filter unwanted outliers. ...
- Step 4: Handle missing data. ...
- Step 5: Validate and QA.

DATASET!

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.metrics import f1_score, accuracy_score, confusion_matrix ,classification_report
from sklearn.ensemble import RandomForestClassifier
```

In [2]:
```
df=pd.read_csv("C:/Users/siddh/dataset.csv")
df.head()
```

Out[2]:

	Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4	Symptom_5	Symptom_6	Symptom_7	Symptom_8	Symptom_9	Symptom_10	Sy
0	Fungal infection	itching	skin_rash	nodal_skin_eruptions	dischromic _patches	NaN	NaN	NaN	NaN	NaN	NaN	
1	Fungal infection	skin_rash	nodal_skin_eruptions	dischromic _patches	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
2	Fungal infection	itching	nodal_skin_eruptions	dischromic _patches	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
3	Fungal infection	itching	skin_rash	dischromic _patches	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
4	Fungal infection	itching	skin_rash	nodal_skin_eruptions	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

In [3]:
```
df.tail()
```

Out[3]:

	Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4	Symptom_5	Symptom_6	Symptom_7	Symptom_8	Sym
4915	(vertigo) Paroymsal Positional Vertigo	vomiting	headache	nausea	spinning_movements	loss_of_balance	unsteadiness	NaN	NaN	
4916	Acne	skin_rash	pus_filled_pimples	blackheads	scurring	NaN	NaN	NaN	NaN	
4917	Urinary tract infection	burning_micturition	bladder_discomfort	foul_smell_of urine	continuous_feel_of_urine	NaN	NaN	NaN	NaN	
4918	Psoriasis	skin_rash	joint_pain	skin_peeling	silver_like_dusting	small_dents_in_nails	inflammatory_nails	NaN	NaN	
4919	Impetigo	skin_rash	high_fever	blister	red_sore_around_nose	yellow_crust_ooze	NaN	NaN	NaN	

In [4]:
```
df_desc=pd.read_csv("C:/Users/siddh/symptom_Description.csv")
df_prec=pd.read_csv("C:/Users/siddh/symptom_precaution.csv")
df_sev=pd.read_csv("C:/Users/siddh/Symptom-severity.csv")
```

In [11]:
```
df=df.replace(np. nan,'Asymtomatic',regex=True)
df.head()
```

Out[11]:

	Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4	Symptom_5	Symptom_6	Symptom_7	Symptom_8	Symptom_9
0	Fungal infection	itching	skin_rash	nodal_skin_eruptions	dischromic _patches	Asymtomatic	Asymtomatic	Asymtomatic	Asymtomatic	Asymtomatic
1	Fungal infection	skin_rash	nodal_skin_eruptions	dischromic _patches	Asymtomatic	Asymtomatic	Asymtomatic	Asymtomatic	Asymtomatic	Asymtomatic
2	Fungal infection	itching	nodal_skin_eruptions	dischromic _patches	Asymtomatic	Asymtomatic	Asymtomatic	Asymtomatic	Asymtomatic	Asymtomatic
3	Fungal infection	itching	skin_rash	dischromic _patches	Asymtomatic	Asymtomatic	Asymtomatic	Asymtomatic	Asymtomatic	Asymtomatic
4	Fungal infection	itching	skin_rash	nodal_skin_eruptions	Asymtomatic	Asymtomatic	Asymtomatic	Asymtomatic	Asymtomatic	Asymtomatic

In [12]:
```
df_prec=df_prec.replace(np. nan,'no precaution',regex=True)
df_prec.head()
```

Out[12]:

	Disease	Precaution_1	Precaution_2	Precaution_3	Precaution_4
0	Drug Reaction	stop irritation	consult nearest hospital	stop taking drug	follow up
1	Malaria	Consult nearest hospital	avoid oily food	avoid non veg food	keep mosquitos out
2	Allergy	apply calamine	cover area with bandage	no precaution	use ice to compress itching
3	Hypothyroidism	reduce stress	exercise	eat healthy	get proper sleep
4	Psoriasis	wash hands with warm soapy water	stop bleeding using pressure	consult doctor	salt baths

Training and testing the data

Naive Bayes

$$P(b \vee a) = \frac{P(a \vee b)P(b)}{P(a)}$$

Above,

- P(b|a) is that the posterior chance of class (b, target) given predictor (a, attributes).
- P (b) is the prior probability of class.
- P(a|c)iis that chance that is that the chance of predictor given class.
- P(a) is the prior probability of predictor.

Training to Testing

70:30

In [46]:
```
x_train, x_test, y_train, y_test = train_test_split(df_data, label, shuffle=True, train_size = 0.70)
randomFC = RandomForestClassifier()
randomFC.fit(x_train, y_train)
result = randomFC.predict(x_test)
print(randomFC)
print(classification_report(y_true=y_test, y_pred=result))
print('F1-score% =', f1_score(y_test, result, average='macro')*100, '|', 'Accuracy% =', accuracy_score(y_test, result)*100)
```

```
RandomForestClassifier()
                                   precision    recall  f1-score   support

(vertigo)_Paroymsal__Positional_Vertigo       1.00      1.00      1.00        39
                                  AIDS       1.00      1.00      1.00        42
                                  Acne       1.00      1.00      1.00        31
                  Alcoholic_hepatitis        1.00      1.00      1.00        26
                               Allergy       0.86      1.00      0.92        36
                             Arthritis       1.00      1.00      1.00        34
                        Bronchial_Asthma       1.00      1.00      1.00        35
                    Cervical_spondylosis       0.94      1.00      0.97        33
                             Chicken_pox       1.00      1.00      1.00        40
                     Chronic_cholestasis       1.00      1.00      1.00        35
                             Common_Cold       1.00      1.00      1.00        30
                                  Dengue       1.00      1.00      1.00        37
                                Diabetes       1.00      1.00      1.00        30
            Dimorphic_hemmorhoids(piles)       1.00      1.00      1.00        38
```

Hepatitis_E	1.00	1.00	1.00	33
Hypertension	0.97	0.95	0.96	38
Hyperthyroidism	1.00	1.00	1.00	36
Hypoglycemia	1.00	1.00	1.00	32
Hypothyroidism	1.00	1.00	1.00	35
Impetigo	1.00	0.98	0.99	42
Jaundice	1.00	1.00	1.00	30
Malaria	1.00	1.00	1.00	40
Migraine	1.00	1.00	1.00	38
Osteoarthristis	1.00	1.00	1.00	34
Paralysis_(brain_hemorrhage)	1.00	0.91	0.96	47
Peptic_ulcer_diseae	1.00	1.00	1.00	36
Pneumonia	1.00	1.00	1.00	30
Psoriasis	1.00	1.00	1.00	39
Tuberculosis	1.00	1.00	1.00	33
Typhoid	1.00	1.00	1.00	42
Urinary_tract_infection	1.00	1.00	1.00	34
Varicose_veins	1.00	1.00	1.00	48
hepatitis_A	1.00	1.00	1.00	32
accuracy			0.99	1476
macro avg	0.99	0.99	0.99	1476
weighted avg	0.99	0.99	0.99	1476

F1-score% = 99.17111399812923 | Accuracy% = 99.11924119241192

CONCLUSION

This project aims to predict the disease on the basis of the symptoms. The project is designed in such a way that the system takes symptoms from the user as input and produces output i.e. predict disease.

In conclusion, for disease risk modelling, the accuracy of risk prediction depends on the diversity feature of the hospital data.