EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

A predictive prototype for the identification of diseases relied on the symptoms described by patients

Suvendu Kumar Nayak[1,*], Mamata Garanayak[2] and Sangram Keshari Swain[3]

[1,3]Centurion University of Technology and Management, Odisha, India
[2]KISS Deemed to be University, Bhubaneswar, Odisha, India

Abstract

INTRODUCTION: A thorough and timely investigation of any health-related problem is essential for disease prevention and treatment. The normal way of diagnosis may not be sufficient in the event of a serious illness problem.
OBJECTIVE: Creating a medical diagnosis prototype that uses many machine learning processes to forecast any illness relied on symptoms explained by patients can lead to an errorless diagnosis as compared to the traditional ways.
METHODS: We created a disease prediction prototype using ML techniques such as random forest, CART, multinomial linear regression, and KNN. The data set utilized for processing contained over 132 illnesses. Diagnosis algorithm outcomes the ailment that the person may be suffering from relied on the symptoms provided by the patients.
RESULTS: When compared to CART and random forest (accuracy is 97.72%, multinomial linear regression and KNN produced the best outcomes. The accuracy of the KNN prediction and multinomial linear regression techniques was 98.76%.
CONCLUSION: The diagnostic prototype can function as a doctor in the early detection of an illness, ensuring that medical care can begin in an appropriate time and many lives can be secured.

Keywords: Prediction of disease, k-means, Random forest, Multinomial linear regression, CART prototype, KNN

Received on 11 December 2024, accepted on 08 March 2024, published on 12 March 2024

doi: 10.4108/eetpht.10.5405

1. Introduction

At the moment, per capita medical supplies are few, and high quality medical supplies are concentrated in major cities and substantial institutions. Even if patient's symptoms are minor, many patients are concerned about their health and go to major hospitals for top medical care. Dispute and constraints among medical supplies availability and stipulation are long standing circumstances. Patients naturally care about the links among symptoms and illnesses during medical consultations. Now a day, most individuals post symptoms online to acquire prediagnosis results, with the goal of screening serious diseases and seeking recommendations for subsequent proper medical treatment [1].

Because of heightening of the power of the computer and the availability of the datasets on open-source sources, ML has expanded in demand as technology has improved. In health care, ML is utilized in a variety of means. The health care business creates a large amount of information in the form of pictures, patient information, and other sorts of information that may be utilised to identify trends and build predictions. ML is utilized in health care to resolve a range of problems [2].

Prediction prototypes are intended to help healthcare providers and people make decisions regarding diagnostic tests, starting or ending treatments, or changing their lifestyle. While not a replacement for clinical experience, they can give objective facts on a person's illness risk and help to avoid certain frequent biases in clinical decision making. Biases in the way data is acquired or filtered for use by the prototype, on the other hand, might introduce other forms of biases, therefore the choice of underlying data and cohort selection are critical. Furthermore, information production in health care is rapidly increasing and outpacing human cognition's ability to manage [3]. Allowing prototypes to influence

[*]Corresponding author. Email: suvendu.sonu@gmail.com

decision-making to support human cognition is a scalable strategy to manage expanding data quantities and information complexity.

An intelligent information prototype that can carry out pre-diagnosis automatically relied on the symptoms supplied by patients can help to ease the issue of medical supplies scarcity. Such diagnostic approaches are provided in this paper. The studies relevant to the problem statement are discussed in the second portion of the article. The technique and intended work are elaborated on in the third part. The results are analysed in the fourth section, and the conclusion is stated in the fifth section, followed by references.

2. Related Work

By combining correlation distance measurements and the k-nearest neighbor algorithm, Singh, A., and Pandey, B. [4] suggested a unique method for identifying liver disorders. The performance of the suggested KNN approach was compared by the authors using a wide range of classifiers (LDA, DLDA, QDA, DQDA, and LSSVM), and it was confirmed that the new approach outperformed others. On testing data, the proposed model's accuracy is very near to 97%. This experiment uses a dataset of 583 samples and 10 input features that were taken from the university of California ML repository.

A systematic evaluation of the literature has been conducted by Parimbelli et al. [5] to investigate the techniques and strategies for determining patient similarity, a crucial component in customizing medical care for each patient. To enhance disease diagnosis, therapy, and management, a total of 279 papers were examined along four dimensions (data forms, clinical domains, data analysis methodologies, and translational stage of study findings).

Sartakhti et al. [6] present a novel hybrid approach that combines SVM (Support Vector Machine) and SA (Simulated Annealing) approaches for the diagnosis of hepatitis disease. The proposed hybrid approach showed how the SA's optimization procedure can improve the SVM's classification abilities. The multivariate dataset for the hepatitis disease acquired from the UCI MLg Repository. The collection has 19 number of attributes and 155 number of samples. SVM-SA attained classification accuracy is 96.25%, which is quite encouraging.

Olsen C et al. [7] evaluated the state of machine learning in the testing, classifying, and forecasting of heart failure. They go over the several datasets used, such as the Multi-Ethnic Study of Atherosclerosis and the Framingham Heart Study, in addition to the different approaches and algorithms used, like decision trees, support vector machines, and deep learning.

To uncover distinctive clinical trends among dengue patients, Macedo Hair et al. [8] explore the creative application of unsupervised machine learning algorithms. The clinical characteristics of 523 confirmed dengue patients were analyzed using self-organizing maps and random forests, two unsupervised machine learning methodologies, to find natural patterns. The results of this study imply that age plays an important role in deciding how dengue manifests clinically.

The utilization of ML approaches to forecast disease risks is thoroughly explored in Saranya and Pravin's [9] paper. The authors discuss the use of methods such as decision trees, support vector machines, neural networks, and ensemble approaches to treat a variety of illnesses. The authors offer helpful insights into the best practices and concerns for academics and practitioners in the field by analyzing the advantages and disadvantages of different strategies.

Razavian and Sontag [10] make an important addition to the field of medical diagnostics by introducing the use of Temporal Convolutional Neural Networks (TCN) for extracting pertinent information from time-series data received from lab tests. The suggested approach uses an individual's lab test results as input to learn to forecast the likelihood that person would get a specific ailment. The study used 298K individuals and eight years' worth of lab test results. The outcomes demonstrated that the suggested method greatly outperformed the benchmark techniques.

Nithya and Ilango [11] go through the many ML algorithms that can be used for predictive analytics as well as the various uses of predictive analytics in the healthcare industry. The authors highlight the positive aspects of employing various ML algorithms and ML tools as they explore various issues in healthcare data, such as diverse, noisy, and incomplete data.

For patients with Parkinson's Disease, Shamir et al. [12] created a machine learning-based model to optimize the combination of deep brain stimulation and drug therapy. The researchers gathered information from PD patients who had DBS surgery and examined a range of clinical and electrophysiological factors, including symptom intensity, medication dosages, and patterns of brain activity. According to the study's findings, the algorithms' accuracy increased when only certain symptoms were considered into account.

A system for real time ECG monitoring and CVD prediction built on a cell phone was created by Jin et al. [13]. ECG signals are classified as normal or abnormal by the system using an adaptive machine learning technique. An ECG signal dataset from patients with and without cardiovascular disease is used to train the algorithm. Ninety percent of the ECG data could be accurately classified as normal or abnormal using the adaptive machine learning method.

A machine learning-based approach for disease prediction based on symptoms is proposed by Deepthi et al. [14] The system makes use of a collection of 4000 records, each containing details about the patient's symptoms, medical background, and diagnosis. The system employs decision trees, naive Bayes, and random forests as its three machine learning algorithms. The findings indicate that the random forest algorithm can accurately forecast diseases up to 94.6% of the time.

A method created by Kanchan and Kishore [15] employs PCA and machine learning algorithms to forecast particular diseases. Using information from the hospital repository, which includes 1856 records, the authors examine the effectiveness of this strategy in the context of a cardiac condition. Decision trees, Naive Bayes, and SVMs were the three machine learning procedures that were evaluated in the article. The outcomes demonstrated that the SVM algorithm performed at its best when combined with PCA.

The ability of classification algorithms to correctly forecast liver illness is demonstrated by Singh et al. [16]. In order to distinguish between those with healthy livers and those who have been diagnosed with liver disease, the study evaluates how well several algorithms, such as logistic regression, k-nearest neighbours, and support vector machines, perform. The authors evaluate the accuracy, sensitivity, and other pertinent metrics of various algorithms for diagnosing liver illness using data from the Indian Liver Patient Dataset (ILPD) dataset. The logistic regression algorithm outperformed the other two algorithms with an accuracy of 73.97%.

Grampurohit and Sagarnal [17] add to the expanding body of work on machine learning-based disease prediction. Utilising the three data mining algorithms—DT Classifier, RF Classifier, and Naive Bayes Classifier—disease prediction system is put into action. A comparison study reveals that each of the three procedures performs on a medical record with the accuracy of up to 95%. The trial involved 132 symptoms and 41 different diseases, and it involved 4920 occurrences.

Hamsagayathri and Vigneshwaran [18] explore the viability of using machine learning approaches for disease prediction based on symptoms. The authors offer insightful information on the relative benefits and drawbacks of various algorithms for handling datasets with symptoms. The study also included a collection of tools created by the AI community.
In order to accurately predict disease, Dahiwade et al. [19] built a model of predictive ability based on the clinical signs of the patients by utilising the KNN and Convolutional Neural Network machine learning procedures. The study compares the KNN and CNN algorithms' performance metrics and finds that CNN outperforms KNN in terms of accuracy and processing speed.

In the research work of Alexander et al. [20], the task of discovering and assessing clinical subgroups of Alzheimer's Disease (AD) using unsupervised machine learning techniques is explored. K-means, kernel k- means, affinity propagation, and latent class analysis are four different clustering techniques that have the potential to reveal underlying patterns and subtypes that conventional clinical assessments could miss. Several evaluative measures revealed that K-means produced the most reliable findings.

The music recommendation system created by Garanayak, M. et al. [21] focuses primarily on content-based, collaborative-based, and popularity-based filtering algorithms while also proposing a hybrid method that combines these three filtering types.

Choudhury, S. et al. [22] developed a movie recommendation system that overcomes the cold start issue, data sparsity, and malicious attack. The authors also propose 4 different recommendation models—Back propagation (BPNN) model, SVD (Singular Value Decomposition) model and DNN with Trust—were contrasted in order to suggest the right movie to the user. The best model, based on the outcomes, has a high accuracy of 83% and a 0.74 MSE value. It is the DNN with trust model.

3. Proposed Work

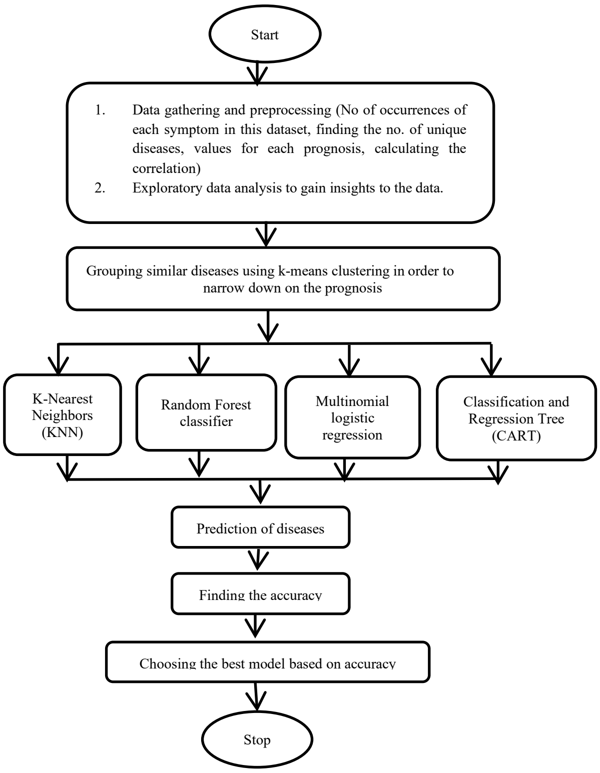
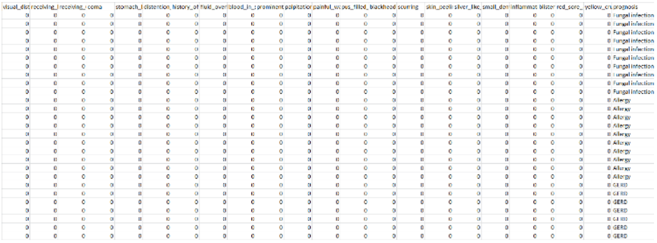

Start
1. Data gathering and preprocessing (No of occurrences of each symptom in this dataset, finding the no. of unique diseases, values for each prognosis, calculating the correlation)
2. Exploratory data analysis to gain insights to the data.
Grouping similar diseases using k-means clustering in order to narrow down on the prognosis
K-Nearest Neighbors (KNN)
Random Forest classifier
Multinomial logistic regression
Classification and Regression Tree (CART)
Prediction of diseases
Finding the accuracy
Choosing the best model based on accuracy
Stop

Figure 1. Proposed workflow

In the Fig. 1. first the dataset is collected and then pre-processed to get the required data set for our work. After getting the required data several predictive models such as k-means clustering, random forest, multinomial logistic regression and CART models are applied to get the prediction outcome and accuracy of each predictive prototype. At last based on the accuracy comparison of several predictive prototypes, the best predictive prototype is chosen.

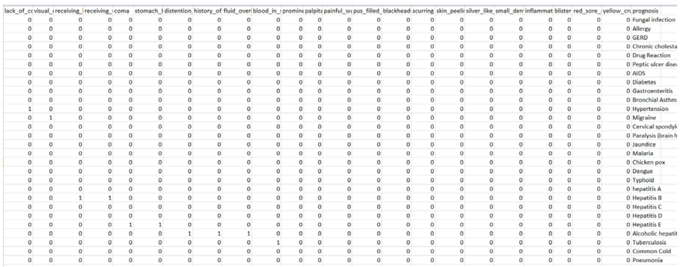
3.1. Data Gathering and Preprocessing

The datasets is collected from the kaggle website which consists of two .csv files; train and test. The train and test files (Fig.2. and Fig.3.) contain 132 features that represent 132 different symptoms and 1 column named “prognosis” that represents the response variable.



visual_dist	receiving_I	receiving_	coma	stomach_I	distention_	history_of	fluid_over	blood_in_:	prominent	palpitatior	painful_w:	pus_filled_	blackhead	scurring	skin_peeli	silver_like_	small_den	inflammat	blister	red_sore_	yellow_cru	prognosis
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	GERD
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	GERD
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	GERD
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	GERD
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	GERD
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	GERD
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	GERD

Figure 2. Train dataset

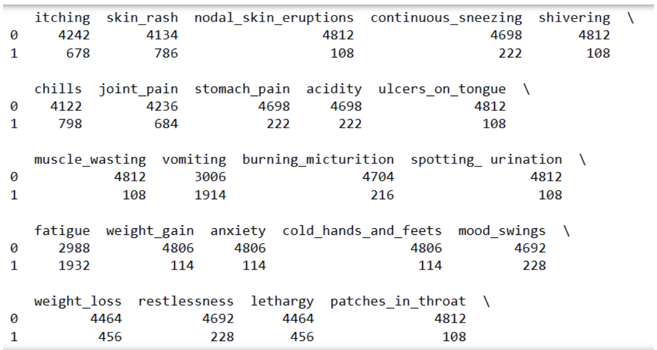


lack_of_cc	visual_(	receiving_I	receiving_(	coma	stomach_I	distention_	history_of	fluid_over	blood_in_:	promine	palpita	painful_wa	pus_filled_	blackhead	scurring	skin_peeli	silver_like_	small_den	inflammat	blister	red_sore_	yellow_cru	prognosis
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	GERD
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Chronic cholesta
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Drug Reaction
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Peptic ulcer dise:
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	AIDS
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Diabetes
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Gastroenteritis
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Bronchial Asthm
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Hypertension
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Migraine
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Cervical spondyl
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Paralysis (brain h
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Jaundice
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Malaria
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Chicken pox
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Dengue
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Typhoid
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	hepatitis A
0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Hepatitis B
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Hepatitis C
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Hepatitis D
0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Hepatitis E
0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Alcoholic hepatit
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	Tuberculosis
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Common Cold
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Pneumonia

Figure 3. Test dataset

Exploratory Data Analysis

In the data analysis section first the number of occurrences of each symptom in this dataset is found (Fig. 4.), then Count the number of occurrences of '1' in each column and sort in descending order (Fig. 5.). After that the number of unique diseases from the prognosis column is 41 as identified and the number of values for each prognosis is found 120 (Fig. 6.).Finally the correlation matrix (Fig.7.) is
calculated to summarize the data, as an input and diagnostic for more advanced analyses.



```
   itching  skin_rash  nodal_skin_eruptions  continuous_sneezing  shivering  \
0     4242       4134                  4812                 4698       4812
1      678        786                   108                  222        108

   chills  joint_pain  stomach_pain  acidity  ulcers_on_tongue  \
0    4122        4236          4698     4698              4812
1     798         684           222      222               108

   muscle_wasting  vomiting  burning_micturition  spotting_ urination  \
0            4812      3006                 4704                 4812
1             108      1914                  216                  108

   fatigue  weight_gain  anxiety  cold_hands_and_feets  mood_swings  \
0     2988         4806     4806                  4806         4692
1     1932          114      114                   114          228

   weight_loss  restlessness  lethargy  patches_in_throat  \
0         4464          4692      4464               4812
1          456           228       456                108
```

Figure 4. Sample of Number of occurrences of each symptom in this dataset

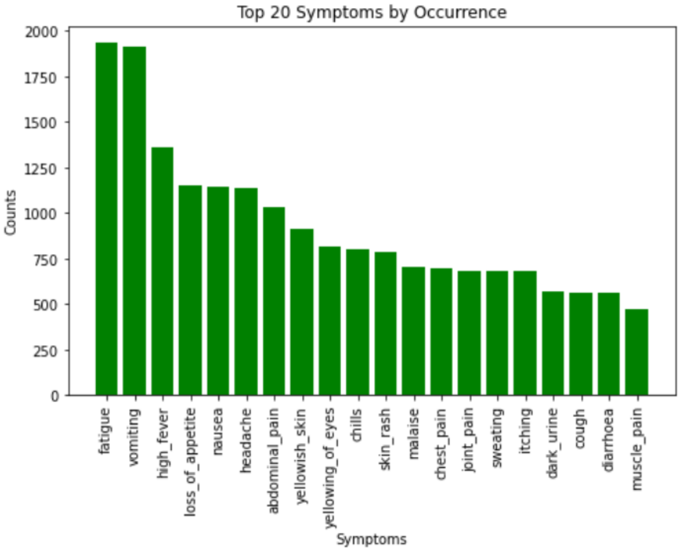


Figure 5. Highest 20 symptoms

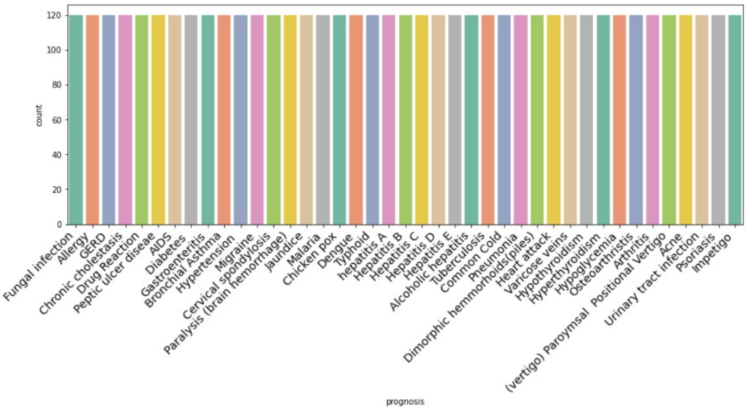


Figure 6. Number of values of each prognosis (120)

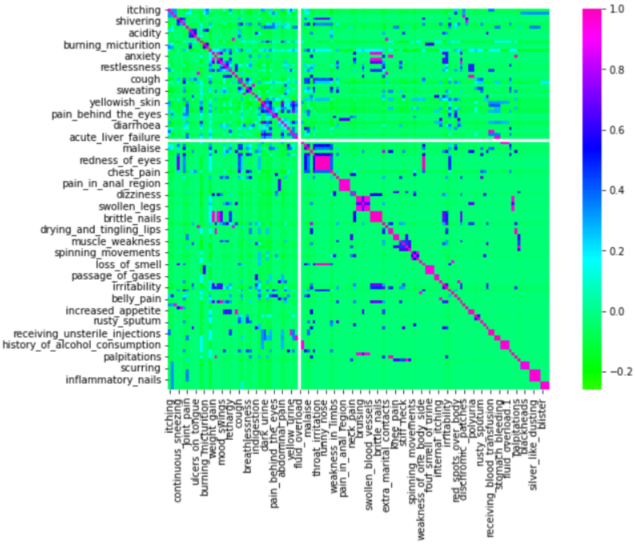


Figure 7. Correlation matrix

3.2. K-means Clustering

One of the clustering procedure is the k means which is a vector quantization approach obtained from signal processing that divide n observations into k number of clusters, with each inspection belongs to the cluster with the closest mean that is cluster centres or centroid, which be of use to as the prototype of the cluster. Here, first the optimal no. of clusters by utilizing silhouette score is determined and then silhouette scores for each number of clusters are plotted (Fig. 8.).

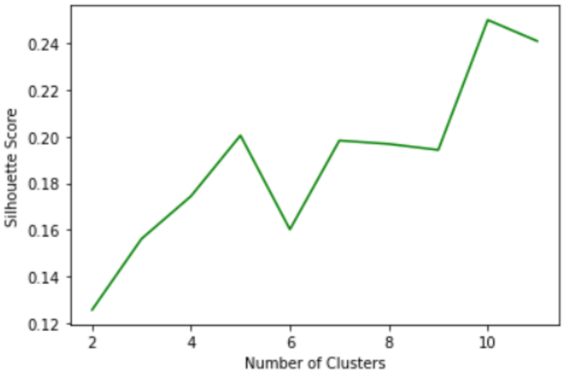
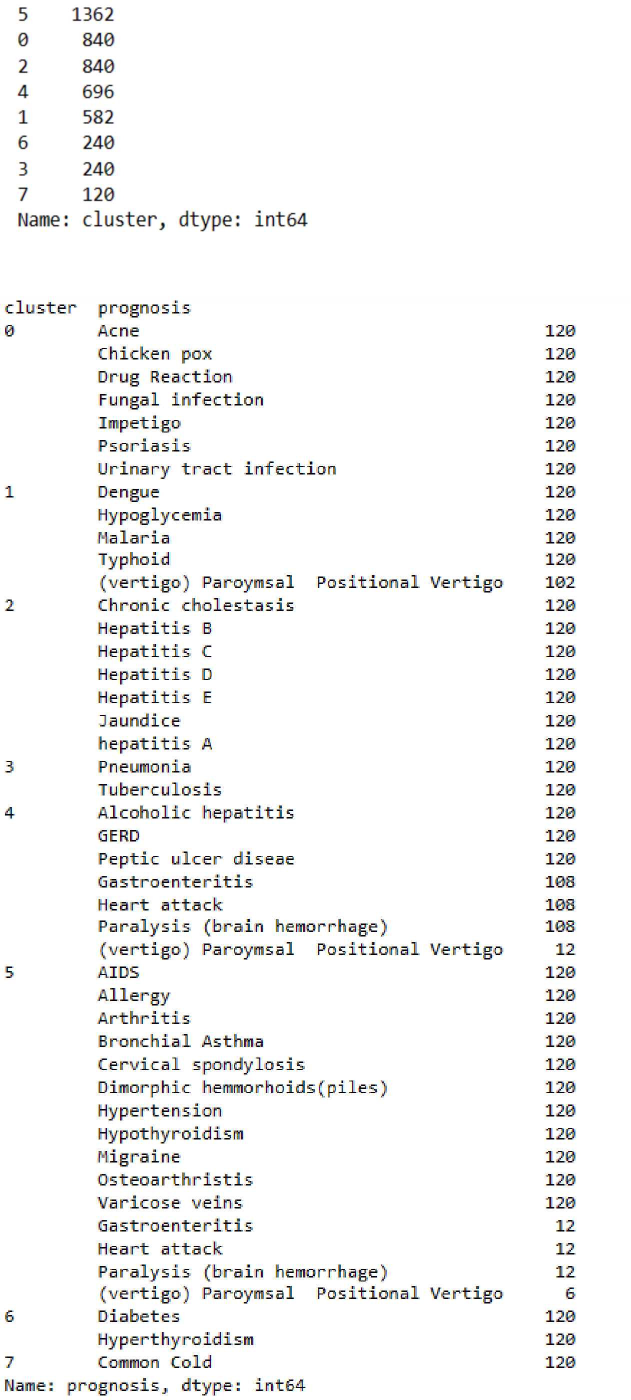


Figure 8. Silhouette scores for each number of clusters

Finally, by choosing the number of clusters with the towering silhouette score performs the clustering (k means) as shown in Fig. 9.

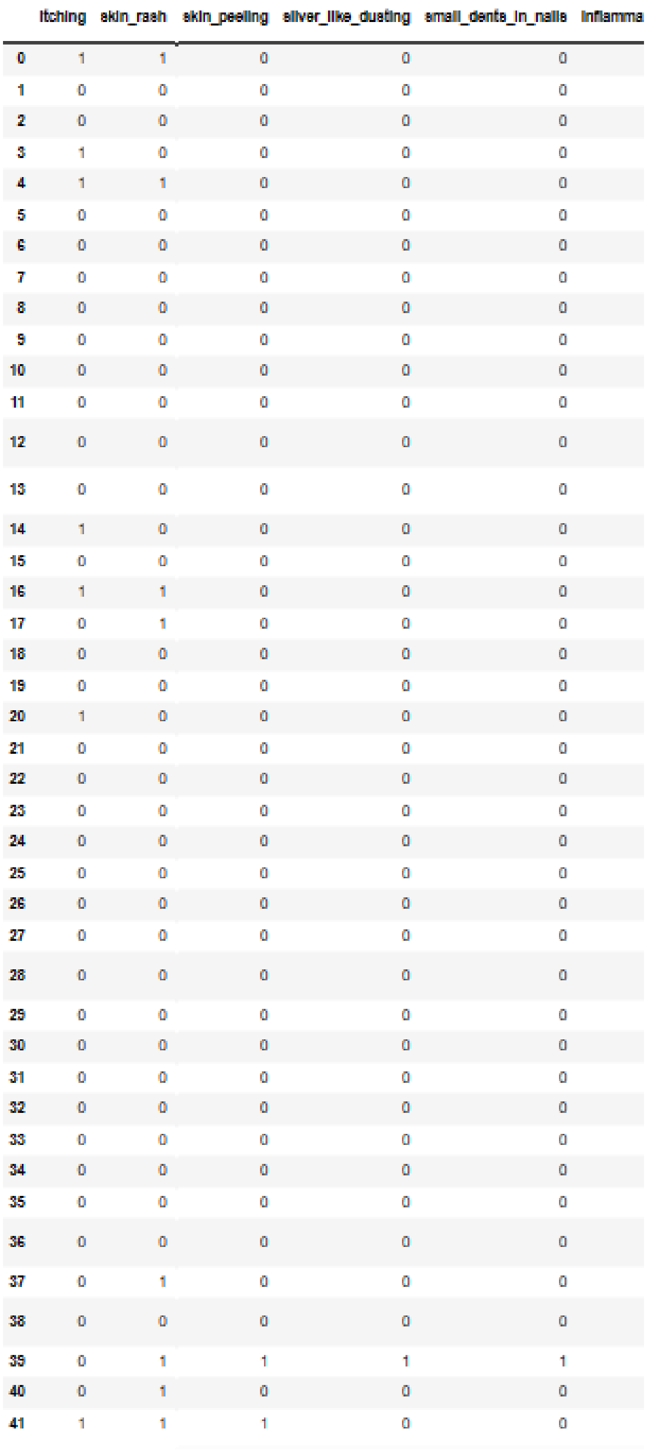


```
5    1362
0     840
2     840
4     696
1     582
6     240
3     240
7     120
Name: cluster, dtype: int64
```

```
cluster  prognosis
0        Acne                                       120
         Chicken pox                                120
         Drug Reaction                              120
         Fungal infection                           120
         Impetigo                                   120
         Psoriasis                                  120
         Urinary tract infection                    120
1        Dengue                                     120
         Hypoglycemia                               120
         Malaria                                    120
         Typhoid                                    120
         (vertigo) Paroymsal  Positional Vertigo    102
2        Chronic cholestasis                        120
         Hepatitis B                                120
         Hepatitis C                                120
         Hepatitis D                                120
         Hepatitis E                                120
         Jaundice                                   120
         hepatitis A                                120
3        Pneumonia                                  120
         Tuberculosis                               120
4        Alcoholic hepatitis                        120
         GERD                                       120
         Peptic ulcer diseae                        120
         Gastroenteritis                            108
         Heart attack                               108
         Paralysis (brain hemorrhage)               108
         (vertigo) Paroymsal  Positional Vertigo     12
5        AIDS                                       120
         Allergy                                    120
         Arthritis                                  120
         Bronchial Asthma                           120
         Cervical spondylosis                       120
         Dimorphic hemmorhoids(piles)               120
         Hypertension                               120
         Hypothyroidism                             120
         Migraine                                   120
         Osteoarthristis                            120
         Varicose veins                             120
         Gastroenteritis                             12
         Heart attack                                12
         Paralysis (brain hemorrhage)                12
         (vertigo) Paroymsal  Positional Vertigo      6
6        Diabetes                                   120
         Hyperthyroidism                            120
7        Common Cold                                120
Name: prognosis, dtype: int64
```

Figure 9. K-means clustering

Fig. 9. depicts that the illnesses are classified according to cluster assignment. This implies that the symptoms of illnesses in the same cluster are likely to be similar. The prediction result of k-means procedure is shown in Fig. 10.



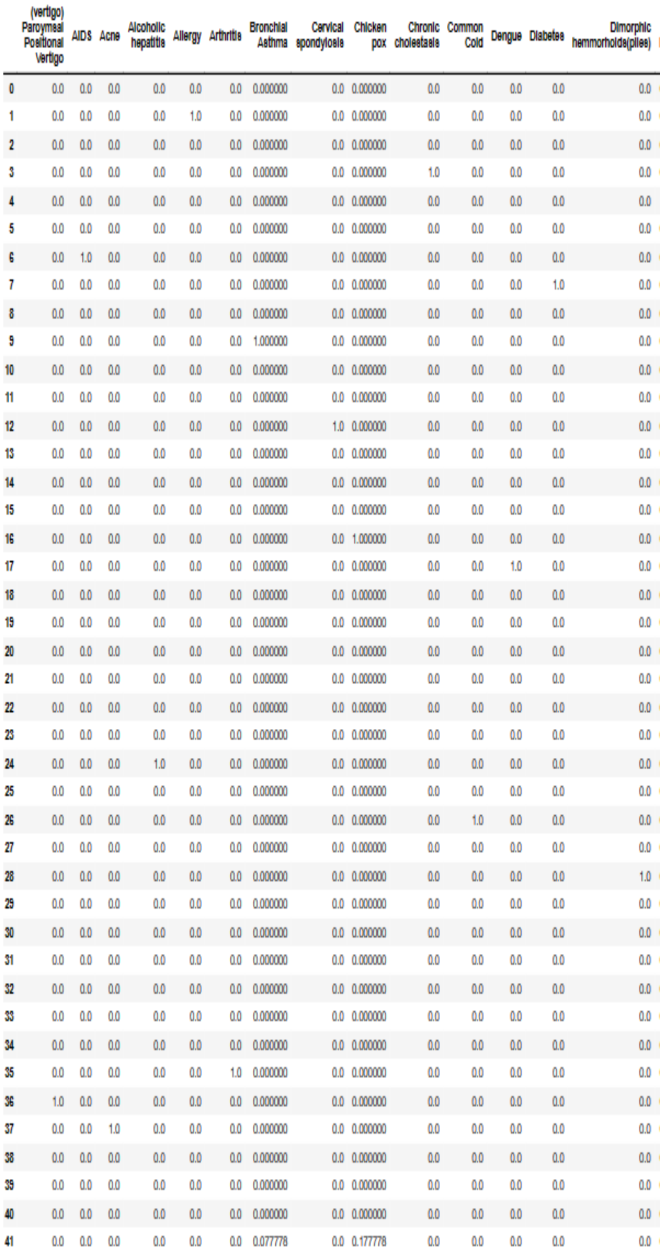
	itching	skin_rash	skin_peeling	silver_like_dusting	small_dents_in_nails	inflamma
0	1	1	0	0	0	
1	0	0	0	0	0	
2	0	0	0	0	0	
3	1	0	0	0	0	
4	1	1	0	0	0	
5	0	0	0	0	0	
6	0	0	0	0	0	
7	0	0	0	0	0	
8	0	0	0	0	0	
9	0	0	0	0	0	
10	0	0	0	0	0	
11	0	0	0	0	0	
12	0	0	0	0	0	
13	0	0	0	0	0	
14	1	0	0	0	0	
15	0	0	0	0	0	
16	1	1	0	0	0	
17	0	1	0	0	0	
18	0	0	0	0	0	
19	0	0	0	0	0	
20	1	0	0	0	0	
21	0	0	0	0	0	
22	0	0	0	0	0	
23	0	0	0	0	0	
24	0	0	0	0	0	
25	0	0	0	0	0	
26	0	0	0	0	0	
27	0	0	0	0	0	
28	0	0	0	0	0	
29	0	0	0	0	0	
30	0	0	0	0	0	
31	0	0	0	0	0	
32	0	0	0	0	0	
33	0	0	0	0	0	
34	0	0	0	0	0	
35	0	0	0	0	0	
36	0	0	0	0	0	
37	0	1	0	0	0	
38	0	0	0	0	0	
39	0	1	1	1	1	
40	0	1	0	0	0	
41	1	1	1	0	0	

Figure 10. Prediction using K-means clustering

3.3. Applying Several Predictive Machine Learning Methods

Random Forest Model

RF is an estimator that secures the service of averaging to heighten predicted accuracy and control over fitting by fitting a no. of DT classifiers on several sub samples of the dataset. If bootstrap=True, then the size of the sub-sample is controlled by the max_samples argument; or else, the whole dataset is utilized to build each tree. Here n_estimators=90 and random_state=40 is utilized to obtain the prediction outcome
(Fig, 11.).

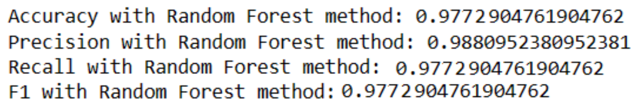


	(vertigo) Paroymsal Positional Vertigo	AIDS	Acne	Alcoholic hepatitis	Allergy	Arthritis	Bronchial Asthma	Cervical spondylosis	Chicken pox	Chronic cholestasis	Common Cold	Dengue	Diabetes	Dimorphic hemmorhoids(piles)
0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	1.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	1.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
6	0.0	1.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	1.0	0.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	1.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
10	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
11	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
12	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	1.0	0.000000	0.0	0.0	0.0	0.0	0.0
13	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
14	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
15	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
16	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	1.000000	0.0	0.0	0.0	0.0	0.0
17	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	1.0	0.0	0.0
18	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
19	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
20	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
21	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
22	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
23	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
24	0.0	0.0	0.0	1.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
25	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
26	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	1.0	0.0	0.0	0.0
27	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
28	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	1.0
29	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
30	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
31	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
32	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
33	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
34	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
35	0.0	0.0	0.0	0.0	0.0	1.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
36	1.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
37	0.0	0.0	1.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
38	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
39	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
40	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
41	0.0	0.0	0.0	0.0	0.0	0.0	0.077778	0.0	0.177778	0.0	0.0	0.0	0.0	0.0

Figure 11. Prediction Probabilities of each Column using Random Forest

Accuracy of Random Forest Prediction

The accuracy score of random forest prediction is found 97.72% that is given in Fig. 12.

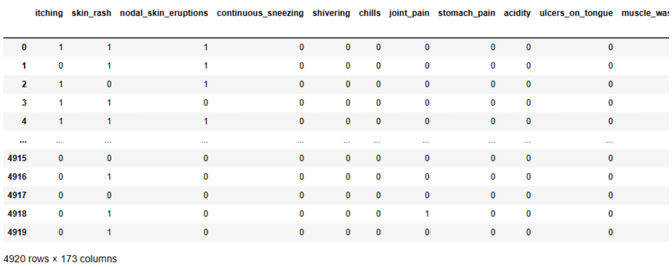


```
Accuracy with Random Forest method: 0.9772904761904762
Precision with Random Forest method: 0.9880952380952381
Recall with Random Forest method: 0.9772904761904762
F1 with Random Forest method: 0.9772904761904762
```

Figure 12. Accuracy Score of the Random Forest Prediction

Multinomial Logistic Regression

Given one or more independent variables, Multinomial Logistic Regression is an approach relied on classification that extends the procedure of logistic regression to tackle multiclass probable outcome issues. This prototype is used to predict the probability of a categorically dependent variable with two or more outcome classes. When the dependent categorical variable has 2 outcome classes, for example, a ward can either "Pass" or "Fail" a test, or a manager of a bank can either "Grant" or "Reject" a loan for a customer, the logistic regression model is used. The prediction probabilities are shown in Fig. 13.



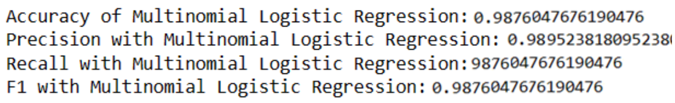
	itching	skin_rash	nodal_skin_eruptions	continuous_sneezing	shivering	chills	joint_pain	stomach_pain	acidity	ulcers_on_tongue	muscle_was
0	1	1	1	0	0	0	0	0	0	0	
1	0	1	1	0	0	0	0	0	0	0	
2	1	0	1	0	0	0	0	0	0	0	
3	1	1	0	0	0	0	0	0	0	0	
4	1	1	1	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	
4915	0	0	0	0	0	0	0	0	0	0	
4916	0	1	0	0	0	0	0	0	0	0	
4917	0	0	0	0	0	0	0	0	0	0	
4918	0	1	0	0	0	0	1	0	0	0	
4919	0	1	0	0	0	0	0	0	0	0	

4920 rows × 173 columns

Figure 13. Prediction Probabilities of each Column using Multinomial Linear Regression

Accuracy of Multinomial Linear Regression

Accuracy score of multinomial linear regression is found 98.76%, which is shown in Fig. 14.



```
Accuracy of Multinomial Logistic Regression: 0.9876047676190476
Precision with Multinomial Logistic Regression: 0.989523818095238
Recall with Multinomial Logistic Regression:9876047676190476
F1 with Multinomial Logistic Regression: 0.9876047676190476
```

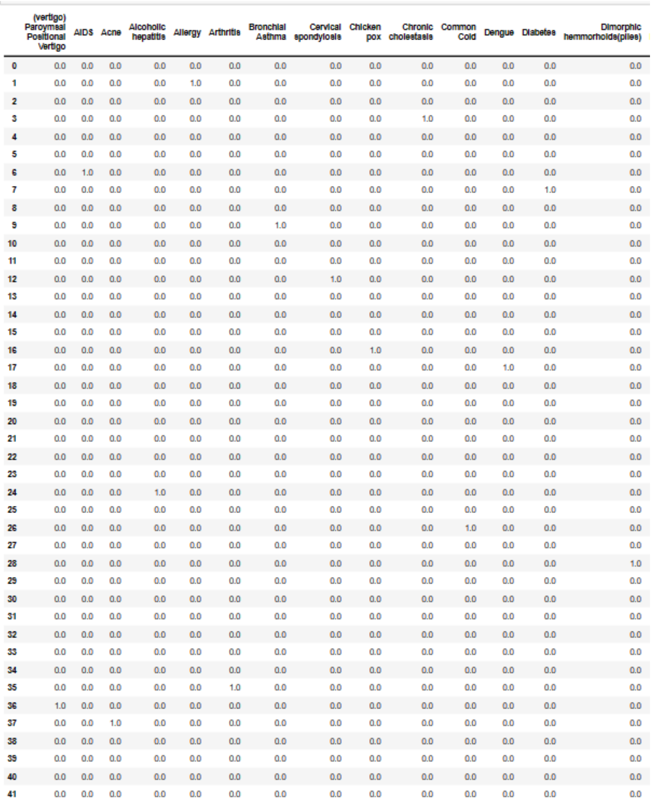
Figure 14. Accuracy Score of Multinomial Linear Regression

CART Prototype

Classification And Regression Tree or CART is a DT procedure variant. It is capable of doing both classification and regression tasks. To train Decision Trees (also known as "growing" trees), Scikit Learn use the Classification and Regression Tree method. The CART procedure operates in the following manner:

i. Determine the appropriate split point for each input.
ii. New "best" split tip is set on relied on the best split tips of each and every input in first step.
iii. Partition the input based on the "best" split tip.
iv. Split until a stopping rule is met or there is no more acceptable splitting available.

Prediction Probabilities of each Column using CART is shown in Fig. 15.

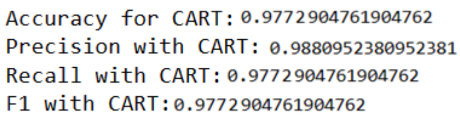


	(vertigo) Paroymsal Positional Vertigo	AIDS	Acne	Alcoholic hepatitis	Allergy	Arthritis	Bronchial Asthma	Cervical spondylosis	Chicken pox	Chronic cholestasis	Common Cold	Dengue	Diabetes	Dimorphic hemmorhoids(piles)
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	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
3	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
4	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
5	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
6	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
7	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
8	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
9	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10	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
11	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
12	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	0.0
13	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
14	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
15	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
16	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
17	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
18	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
19	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
20	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
21	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
22	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
23	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
24	0.0	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
25	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
26	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
27	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
28	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
29	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
30	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
31	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
32	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
33	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
34	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
35	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
36	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
37	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
38	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
39	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
40	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
41	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

Figure 15. Prediction Probabilities of each Column using CART

Accuracy of CART Prediction

The accuracy score of CART prediction is found 97.72%, which is shown in Fig. 16.

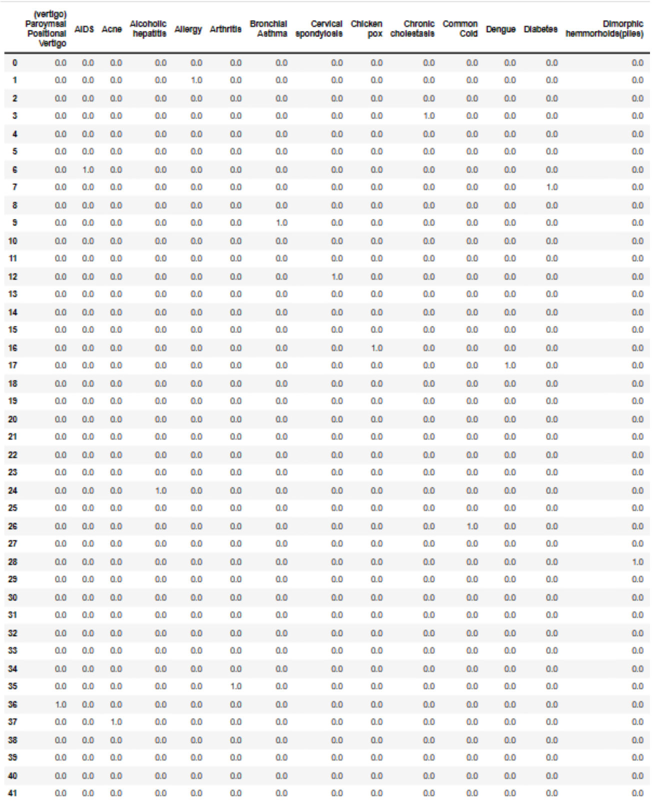


```
Accuracy for CART: 0.9772904761904762
Precision with CART: 0.9880952380952381
Recall with CART: 0.9772904761904762
F1 with CART: 0.9772904761904762
```

Figure 16. Accuracy Score of CART Prediction

KNN Algorithm

KNN procedure utilizes the full training dataset as a reference throughout the phase of training. When producing forecasting, it uses a distance metric such as Euclidean distance to determine the distance between the input data point and all of the training samples. The programme then calculates the distances among the input data point and its K closest neighbours. In the case of classification, the method assigns the most prevalent class label among the K neighbours to the input data point as the predicted label.

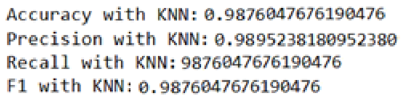


	(vertigo) Paroymsal Positional Vertigo	AIDS	Acne	Alcoholic hepatitis	Allergy	Arthritis	Bronchial Asthma	Cervical spondylosis	Chicken pox	Chronic cholestasis	Common Cold	Dengue	Diabetes	Dimorphic hemmorhoids(piles)
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	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
3	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
4	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
5	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
6	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
7	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
8	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
9	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10	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
11	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
12	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	0.0
13	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
14	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
15	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
16	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
17	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
18	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
19	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
20	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
21	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
22	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
23	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
24	0.0	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
25	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
26	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
27	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
28	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
29	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
30	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
31	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
32	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
33	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
34	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
35	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
36	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
37	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
38	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
39	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
40	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
41	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

Figure 17. Prediction Probabilities of each Column using KNN

Accuracy of KNN

The accuracy score of KNN is found 98.76%, which is shown in Fig. 18.



```
Accuracy with KNN: 0.9876047676190476
Precision with KNN: 0.9895238180952380
Recall with KNN: 9876047676190476
F1 with KNN: 0.9876047676190476
```

Figure 18. Accuracy Score of KNN

Accuracy Comparison

By Utilizing several prediction model such as random forest, multinomial logistic regression, CART method and KNN, it is found that the prediction outcome of random forest and CART model is 97.72% and the prediction outcome of multinomial logistic regression and KNN is 98.76% as given in Table 1.

Table 1. The Accuracy Comparison

Sl. No.	Prototype	Accuracy (%)
1	Random Forest	97.72%
2	Multinomial Logistic Regression	98.76%
3	CART Model	97.72%
4	KNN	98.76%

4. Conclusion

This paper is relied on prediction of disease by utilizing various prototypes of machine learning like multinomial linear regression, random forest, KNN and CART. The prediction accuracy of random forest is 97.72%, multinomial linear regression is 98.76, CART model is 97.72% and KNN is 98.76%. Out of all the accuracies, the multinomial linear regression and KNN shows the better outcomes as comparison to random forest and CART prototype. This prototype for disease diagnosis can be utilized as the doctor of a health issue so that treatment can start on time and many lives can be extricated.

References

[1] Jia, Z., Zeng, X., Duan, H., Lu, X., & Li, H. A patient-similarity-based model for diagnostic prediction. International journal of medical informatics, 135, 104073 (2020).

[2] Lynch, C. J., & Liston, C. New machine-learning technologies for computer-aided diagnosis. Nature medicine, 24(9), 1304-1305 (2018).

[3] Hashem, A. M., Rasmy, M. E. M., Wahba, K. M., & Shaker, O. G. Prediction of the degree of liver fibrosis using different pattern recognition techniques. In 2010 5th Cairo International Biomedical Engineering Conference (pp. 210-214). IEEE (2010, December).

[4] Singh, A., & Pandey, B. (2016). Diagnosis of liver disease using correlation distance metric based k-nearest neighbor approach. In Intelligent Systems Technologies and Applications 2016 (pp. 845-856). Springer International Publishing

[5] Parimbelli, E., Marini, S., Sacchi, L., & Bellazzi, R. (2018). Patient similarity for precision medicine: A systematic review. Journal of biomedical informatics, 83, 87-96.

[6] Sartakhti, J. S., Zangooei, M. H., & Mozafari, K. (2012). Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA). Computer methods and programs in biomedicine, 108(2), 570-579..

[7] Olsen, C. R., Mentz, R. J., Anstrom, K. J., Page, D., & Patel, P. A. (2020). Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure. American Heart Journal, 229, 1-17.

[8] Macedo Hair, G., Fonseca Nobre, F., & Brasil, P. (2019). Characterization of clinical patterns of dengue patients using an unsupervised machine learning approach. BMC infectious diseases, 19, 1-11.

[9] Saranya, G., & Pravin, A. (2020). A comprehensive study on disease risk predictions in machine learning. International Journal of Electrical and Computer Engineering, 10(4), 4217.

[10] Razavian, N., & Sontag, D. (2015). Temporal convolutional neural networks for diagnosis from lab tests. arXiv preprint arXiv:1511.07938.

[11] Nithya, B., & Ilango, V. (2017, June). Predictive analytics in health care using machine learning tools and techniques. In 2017 International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 492-499). IEEE.

[12] Shamir, R. R., Dolber, T., Noecker, A. M., Walter, B. L., & McIntyre, C. C. (2015). Machine learning approach to optimizing combined stimulation and medication therapies for Parkinson's disease. Brain stimulation, 8(6), 1025-1032.

[13] Jin, Z., Sun, Y., & Cheng, A. C. (2009, September). Predicting cardiovascular disease from real-time electrocardiographic monitoring: An adaptive machine learning approach on a cell phone. In 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 6889-6892). IEEE.

[14] Deepthi, Y., Kalyan, K. P., Vyas, M., Radhika, K., Babu, D. K., & Krishna Rao, N. V. (2020). Disease prediction based on symptoms using machine learning. In Energy Systems, Drives and Automations: Proceedings of ESDA 2019 (pp. 561-569). Singapore: Springer Singapore.

[15] Kanchan, B. D., & Kishor, M. M. (2016, December). Study of machine learning algorithms for special disease prediction using principal of component analysis. In 2016 international conference on global trends in signal processing, information computing and communication (ICGTSPICC) (pp. 5-10). IEEE.

[16] Singh, A. S., Irfan, M., & Chowdhury, A. (2018, December). Prediction of liver disease using classification algorithms. In 2018 4th international conference on computing communication and automation (ICCCA) (pp. 1-3). IEEE

[17] Grampurohit, S., & Sagarnal, C. (2020, June). Disease prediction using machine learning algorithms. In 2020 International Conference for Emerging Technology (INCET) (pp. 1-7). IEEE.

[18] Hamsagayathri, P., & Vigneshwaran, S. (2021, February). Symptoms based disease prediction using machine learning techniques. In 2021 Third international conference on intelligent communication technologies and virtual mobile networks (ICICV) (pp. 747-752). IEEE.

[19] Dahiwade, Dhiraj, Gajanan Patle, and Ektaa Meshram. "Designing disease prediction model using machine learning approach." In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pp. 1211-1215. IEEE, 2019.

[20] Alexander, N., Alexander, D. C., Barkhof, F., & Denaxas, S. (2021). Identifying and evaluating clinical subtypes of Alzheimer’s disease in care electronic health records using unsupervised machine learning. BMC Medical Informatics and Decision Making, 21(1), 1-13.

[21] Garanayak, M., Nayak, S. K., Sangeetha, K., Choudhury, T., & Shitharth, S. (2022). Content and Popularity-Based Music Recommendation System. International Journal of Information System Modeling and Design (IJISMD), 13(7), 1-14.

[22] Choudhury, S. S., Mohanty, S. N., & Jagadev, A. K. (2021). Multimodal trust based recommender system with machine learning approaches for movie recommendation. International Journal of Information Technology, 13, 475-482.