Research Result

Disease Prediction using Logistic Regression, Decision Tree Classifier, Random Forest Classifier, K-Nearest Neighbour, Support Vector Classifier

Suraj Sharma[1], Riya Pal[2], Shubham Gupta[3], Tisha Gaur[4], Archana Tomar[5]

[1,2,3,4]Department of IT, ITM Gwalior (M.P.), INDIA
[5]Department of CSE, ITM Gwalior (M.P.), INDIA

ABSTRACT

Preventing and treating illnesses requires accurate and timely analysis of health-related problems, for this we developed a system called "Disease Prediction Model", which focuses on prediction of Diseases. Machine learning algorithms enable these predictions, to predict any disease can lead to a more accurate diagnosis than conventional methods. This Model systems give an output that indicates the disease that the individual might have.

KEYWORDS

Decision Tree Classification, Random Forest Classifier and Support Vector Machine Method and K-Nearest Neighbour Method, these are used to Predict the diseases

1. INTRODUCTION

A primary concern of humanity is healthcare. The WHO considers maintaining proper health to be a fundamental right. In order to maintain one's health, it is important to have access to appropriate health care services. So our model are for disease prediction by the use of this, They can assist medical practitioners in taking early decisions and improving health care. The classification is intended to help physicians. Many people are suffering from diseases, The use of computer technology and machine learning techniques is increasing in the development of medical aid software to assist in early disease diagnosis.

Although many efforts have been made to predict diseases using machine learning algorithms, this is an additional effort. Algorithms like Decision Tree Classification, Random Forest Classifier and Support Vector Machine Method and K-Nearest Neighbour Method, As a result, these can give a remedy to the problem.

2. LITERATURE REVIEW

A lot of research has been conducted on predicting diseases using machine learning algorithms based on symptoms an individual displays.

Shadab Adam Pattekari and Asma Parveen[5], An algorithm for predicting heart diseases was developed by them using the Decision Tree Algorithm, A comparison between a patient's data and a list of qualified values is performed. As a Consequent of this study, patients were able to know their heart related issues by providing some basic information.

Nisha Banu and B. Gomathy [4], Using medical data mining techniques, they analysed different types of heart-related problems. A decision tree shows all possible decisions and their outcomes. The most effective result is achieved by devising a variety of rules. Study criteria included age, sex, smoking, being overweight, drinking alcohol, sugar, heart rate, and blood pressure.

Monto et al. [1] created a statistical model that could predict whether or not a patient had influenza. The study included 3744 adults without vaccinations and Influenza patients who had at least 2 other symptoms of influenza and fever. Out of 3744, Upon laboratory confirmation, 2470 had influenza. Based on this data, They reported a 79% accuracy rate for their model.

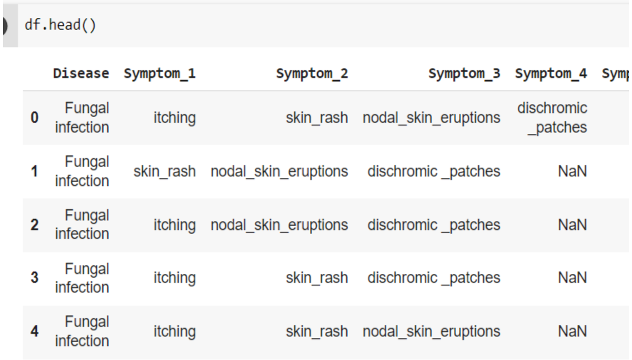
M. Maniruzzaman [3] Based on ML algorithms, they classified diabetes disease. In order to identify diabetes risk factors, logistic regression was used. A 90.62% accuracy was achieved by this Machine Learning based system.

Karayilan et al.[2], By using artificial neural networks in a backpropagation algorithm, they proposed a system for predicting heart diseases. Inputs for the neural network were 13 clinical features and then, With the backpropagation algorithm, neural networks were trained to predict absence or presence of heart diseases. It has an accuracy rate of 95% for predicting these diseases.

3. PROPOSED METHODOLOGY

During execution of evaluation here follows some steps like load data, split data and apply model on it etc for that here 1) Dataset: dataset has 18 columns or features out of which 1 is disease and 17 columns represents the symptoms which are 132 in numbers which are shown in figure 2.

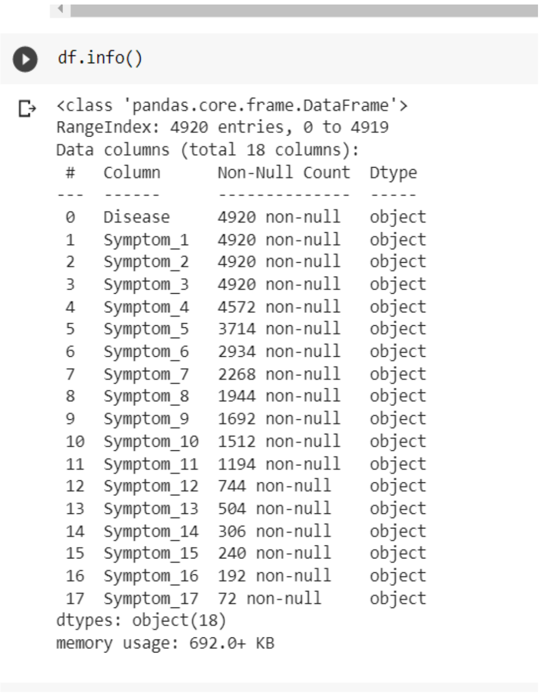
On the basis of these symptoms, a total 42 types of diseases can be predicted which are shown in figure 3.



```
df.head()
```

	Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4	Symp
0	Fungal infection	itching	skin_rash	nodal_skin_eruptions	dischromic _patches	
1	Fungal infection	skin_rash	nodal_skin_eruptions	dischromic _patches	NaN	
2	Fungal infection	itching	nodal_skin_eruptions	dischromic _patches	NaN	
3	Fungal infection	itching	skin_rash	dischromic _patches	NaN	
4	Fungal infection	itching	skin_rash	nodal_skin_eruptions	NaN	

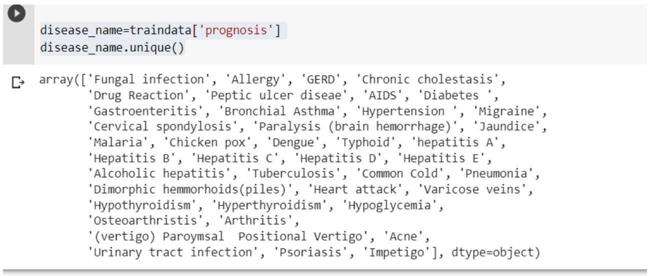
Figure 1 Data set description



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4920 entries, 0 to 4919
Data columns (total 18 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Disease     4920 non-null   object
 1   Symptom_1   4920 non-null   object
 2   Symptom_2   4920 non-null   object
 3   Symptom_3   4920 non-null   object
 4   Symptom_4   4572 non-null   object
 5   Symptom_5   3714 non-null   object
 6   Symptom_6   2934 non-null   object
 7   Symptom_7   2268 non-null   object
 8   Symptom_8   1944 non-null   object
 9   Symptom_9   1692 non-null   object
 10  Symptom_10  1512 non-null   object
 11  Symptom_11  1194 non-null   object
 12  Symptom_12  744 non-null    object
 13  Symptom_13  504 non-null    object
 14  Symptom_14  306 non-null    object
 15  Symptom_15  240 non-null    object
 16  Symptom_16  192 non-null    object
 17  Symptom_17  72 non-null     object
dtypes: object(18)
memory usage: 692.0+ KB
```

Figure2 number of features



```
disease_name=traindata['prognosis']
disease_name.unique()
```

```
array(['Fungal infection', 'Allergy', 'GERD', 'Chronic cholestasis',
       'Drug Reaction', 'Peptic ulcer diseae', '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',
       'Osteoarthristis', 'Arthritis',
       '(vertigo) Paroymsal  Positional Vertigo', 'Acne',
       'Urinary tract infection', 'Psoriasis', 'Impetigo'], dtype=object)
```

Figure 3: Description of diseases

2) Training Data: The training dataset contributes to the accuracy of predictions and performance of the machine learning model.

3) Testing Data: Testing Datasets are used to evaluate the final model objectively.

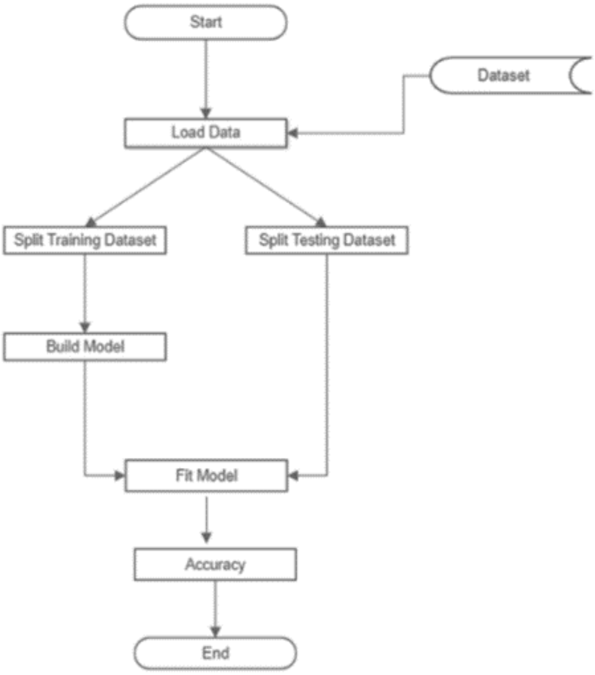


Figure 4: Flow chart of model

In this model, start from splitting the training and testing dataset into input (consisting of symptoms) and the output (as the diseases based on the input factors), After that our next step is to build our model using training data, following the algorithm test the testing dataset and predict the values, resulting in the accuracy of different ML algorithms we use:

1. Logistic Regression

Logistic regression refers to a statistical analysis method that predicts a binary outcome based on prior observations. Using logistic regression models, one or more independent variables are analysed to predict a dependent variable

The underlying technique of Logistic Regression is nearly the same as Linear Regression. In this type of classification, the Logit function is used.

2. Decision Tree

There are several different types of supervised learning algorithms, but the decision tree algorithm is one of them. To make predictions, the decision tree uses a tree diagram at the top. A root node is contained in this feature, which is then split into dominant input features. As long as all input is placed and the node is up and running, this process will continue. In the final node, weights are used to classify inputs.

3. Random Forest Classifier

Random forest classifiers rely on decision trees as their building blocks. In the decision tree, each node represents one measure connected to one characteristic. We compared the results with those obtained from a decision tree.

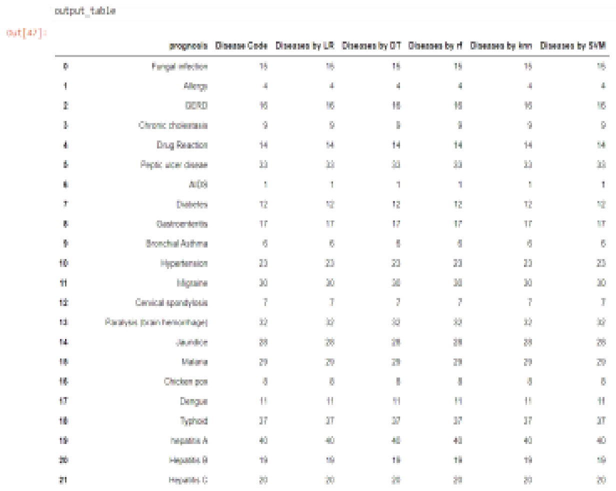
4. K-Nearest Neighbor classifier

The K-nearest neighbors (KNN) algorithm is a type of supervised machine learning algorithm. The distance between a new data point and all other training data points was calculated. The distance can be of Euclidean or Manhattan type. After this, it selects the K nearest data points, K is an integer that can be any value. Lastly, it assigns the data point to the class to which the majority of K data points. In scikit-learn, each neighbour can either be given uniform weight or can be given weights proportional to the inverse of the distance from the query point or a user defined weight.

5. Support Vector Classifier

In machine learning, support vector machines can be used to solve problems related to classification and regression. The majority of its applications, however, are for resolving classification issues. Data distribution and inter-dependencies are not considered in SVM when solving classification problems. It is difficult to identify underlying biological relationships among risk factors due to a limited sample size.

We stored all the predicted value into a table and here the linearity of graph shows the accuracy of algorithm.



```
output_table
```

out[47]:

	prognosis	Disease Code	Diseases by LR	Diseases by DT	Diseases by rf	Diseases by knn	Diseases by SVM
0	Fungal infection	15	15	15	15	15	15
1	Allergy	4	4	4	4	4	4
2	GERD	16	16	16	16	16	16
3	Chronic cholestasis	9	9	9	9	9	9
4	Drug Reaction	14	14	14	14	14	14
5	Peptic ulcer disease	33	33	33	33	33	33
6	AIDS	1	1	1	1	1	1
7	Diabetes	12	12	12	12	12	12
8	Gastroenteritis	17	17	17	17	17	17
9	Bronchial Asthma	6	6	6	6	6	6
10	Hypertension	23	23	23	23	23	23
11	Migraine	30	30	30	30	30	30
12	Cervical spondylosis	7	7	7	7	7	7
13	Paralysis (brain hemorrhage)	32	32	32	32	32	32
14	Jaundice	28	28	28	28	28	28
15	Malaria	29	29	29	29	29	29
16	Chicken pox	8	8	8	8	8	8
17	Dengue	11	11	11	11	11	11
18	Typhoid	37	37	37	37	37	37
19	hepatitis A	40	40	40	40	40	40
20	Hepatitis B	19	19	19	19	19	19
21	Hepatitis C	20	20	20	20	20	20

Figure 5. predicted data table

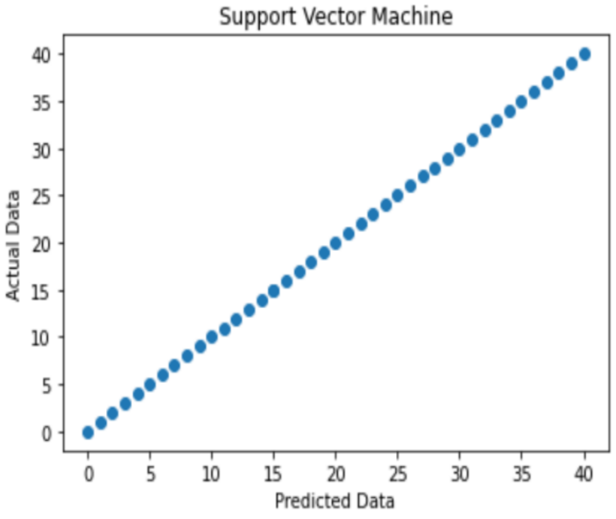


Figure 6. Graph of SVM method

4. RESULTS

We examined the prediction of disease for available input datasets using different machine learning models. We use 5 Machine Learning algorithms out of which 3 gives us 100% accuracy and the other 2 is around 97%.

Table 1: Result of algorithm accuracy

S.No	Algorithm	Accuracy Score
1	Logistic Regression	100%
2	Decision Tree Classifier	97.60%
3	Random Forest Classifier	97%
4	K-Nearest Neighbour	100%
5	Support Vector Classifier	100%

5. CONCLUSION

There is maximum accuracy in almost all the ML models, as some models were parameter dependent, so accuracy score may be variate. We could easily manage the medicine resources required for treating the disease once it has been predicted. With this model, we will be able to lower the costs of treating the disease and we will also be able to improve recovery times. On average, 98% accuracy is achieved in predictions. A successful integration of Disease Predictor was achieved using the system.

REFERENCES

[1] A.S. Monto, S. Gravenstein, M. Elliott, M. Colopy, J. Schweinle, Clinical signs and symptoms predicting influenza infection, Archives of internal medicine 160(21), 3243 (2000)

[2] T. Karayılan, O. Kılı¸c, International Conference on Computer Science and Engineering (UBMK) (IEEE, 2017), pp. 719–723

[3] Nisha Banu, MA; Gomathy, B;. Disease Predicting System Using Data Mining Techniques, (2013).

[4] M. Maniruzzaman, M.J. Rahman, B. Ahammed, M.M. Abedin, Classification and prediction of diabetes disease using machine learning paradigm, Health Information Science and Systems 8(1), 7 (2020)

[5] Adam, S., & Parveen, A.,Prediction System For Heart Disease Using Decision Tree,(2012).

[6] Danaei G, Finucane MM, Lu Y, Singh GM, Cowan MJ, Paciorek CJ, Rao M. National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2.7 million participants. Lancet. 2011;378(9785):31–40.

[7] Luo G. Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction. Health Inf Sci Syst. 2016;4(1):2.

[8] Deniz E, Şengür A, Kadiroğlu Z, Guo Y, Bajaj V, Budak Ü. Transfer learning based histopathological image classification for breast cancer detection. Health Inf Sci Syst. 2018;6(1):18.

[9] Nathan DM. Long-term complications of diabetes mellitus. N Engl J Med. 1993;328(23):1676–85.

[10] Sarwar N, Gao P, Seshasai SR. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease. Lancet. 2010;375(9733):2215–22.

[11] Zimmet P, Alberti KG, Magliano DJ, Bennett PH. Diabetes mellitus statistics on prevalence and mortality: facts and fallacies. Nat Rev Endocrinol. 2016;12(10):616.

[12] Bharath C, Saravanan N, Venkatalakshmi S. Assessment of knowledge related to diabetes mellitus among patients attending a dental college in Salem city—a cross sectional study. Braz Dental Sci. 2017;20(3):93–100.