WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Disease Prediction using Naïve Bayes

Kovvuri Venkata Rama Durgaprasad Reddy[1], Padala Venkata Reddy[2], Geddam Koushik[3], Mrs. T. Sudha Rani4

[1,2,3]Department of CSE, Aditya Engineering College, Surampalem, A.P., India
[4]M. Tech (Ph. D), Associate Professor, Department of CSE, Aditya Engineering College, Surampalem, A.P., India

ABSTRACT:

This system is helps in predicting the system based on the symptoms given by the user to know the most related disease that the user is suffering with. This system is used to predict the disease of the user while he/she is away from the hospital and needs to know what kind of disease it is. Here the system uses the symptoms of the user and processes that information to predict the disease by which the user is affected with. This system uses the Naïve Bayes algorithm mainly to predict the disease based on the symptoms which gives better accuracy than the other systems that are existing now and the result will be more accurate as compared to the existing systems. There is also a comparison among some other algorithms.

Keywords *– System of Prediction; Decision Tree, Naïve Bayes, Random Forest; Disease Prediction; Symptoms; Classification, CNN, KNN;*

I. INTRODUCTION

Now a days, there are many problems that are arising in the health industry such that the inaccurate results of the devices or systems being produced such that they may lead to unaccepted results. In medical

industry, the wrong prediction may lead to a much greater error so that we need to reduce the inaccuracy to the fullest level that we can. The mistakes are not accepted in this industry so the results must be more accurate. As this system is based on Naïve Bayes which has a better accuracy with the classification and it does not require more running time and it can even work on devices with lesser processing capability. This system will give the most accurate prediction so that the user can take the precautionary actions to avoid the aggressiveness of the disease. This system will be useful when the patient is out of range for doctor or the doctor may be on leave or any other restriction caused to the patient to go out such as lockdown etc. At those times, it will be hard to determine the disease caused to the user such that there will be no medical knowledge for some people and to face this kind of challenges, we are introducing this system which uses some intelligent machine learning algorithms to face this challenge which gives more accurate results mostly with a better accuracy than the previous ones. This system is also helpful for the doctors to analyse the pattern of symptoms and the diseases that occur frequently based on the inputs given by the patient. To give better accurate results, the system needs to be trained with different kind of data so that it will be differently trained with different divisions of data and checks the better accuracy to get better testing accuracy. Currently there are some systems that can predict the diseases based on the symptoms such that the data might be complex to handle or it is computationally hard to make the system work which requires more computation and also has to work with large data which may not be handy to work with. So, we need a system that can run on any device with less computational power as much as possible. This system helps the patient to know by which disease he/she is suffering from and helps in taking necessary precautions to reduce the aggression of disease.

II. LITERATURE REVIEW

Designing Disease Prediction Model Using Machine Learning Approach:

Now-a-days, there are many diseases caused to the people because of various living habits and also for environmental conditions such as pollution and other conditions. It is also becoming difficult to determine the diseases based on the symptoms for the doctor too so that the machine learning algorithms are used to determine the disease based on the symptoms. The prediction is based on the machine learning techniques and the data mining is also used for the collection of data. There is a huge amount of medical data such that it can be used to find the pattern in the symptoms and diseases from the early patient. Based on this, we can predict the disease based on the symptoms and the huge data available from the hospitals. This system uses the CNN algorithm to predict the disease based ono the symptoms and it has a better accuracy than the previous model based on KNN and also takes lesser time and space than KNN. The accuracy is better as compared to the previous models.

Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques:

Heart diseases are hard to find as the cardiovascular diseases require novel methods to predict. The prediction can be done based on huge datasets that are produced from the healthcare industry. The ML has been growing in different fields such as Internet of Things (IoT). There are several researches that only give a glimpse of the process to predict the cardiovascular diseases. But in this system, the novel approach is designed to predict the cardiovascular diseases and also the accuracy of the model can be increased much better by using this approach. The prediction model is a combination of several classification techniques and some other feature to predict the disease with higher accuracy. The accuracy is high as compared to other models in this area and the algorithm used is Hybrid Random Forest with a linear model.

Symptoms Based Disease Prediction Using Machine Learning Techniques:

Computer Aided Diagnosis (CAD) is being evolved quickly such that the medical analysis is also being done by the computer but this require correct results. Several CAD applications has been created but most of them are deceptive such that there may be failures and the medial therapies can be different for different diseases but because of the mistake in prediction they may be interchanged. The body organs cannot be determined by a simple equation. Hence, the Machine Learning is introduced here to recognise the patterns and to learn continuously based on the previous disease data. This requires training a system which can be achieved in Machine Learning and hence the disease can be predicted correctly by using the medical analysis data by using algorithms and some decision-making process.

Comparing different supervised machine learning algorithms for disease prediction:

There are several supervised machine learning algorithms and those were compared by taking some previous papers and taking the relevant algorithms used in several systems that are used for the disease prediction models. For this disease prediction models, the supervised machine learning models tend to give better accurate results as compared to the others by feeding them on new data. It has also been observed that the combining of more than one machine learning algorithm has been done to get better results such that there may be different datasets and the accuracy may be different but by combining more than one algorithm can handle multiple cases to give better prediction results than that of given by one algorithm in multiple cases.

Prediction of Heart Disease Using Machine Learning:

The heart strokes have increased in recent times so that we need to keep a system to check the symptoms for a heart stroke such that to prevent from getting one. It is not possible for a common man to undergo the tests like ECG frequently as they are costly to take and also it takes much time to do so. Hence we need a system which is both reliable and time saving for a commoner to predict the chances of a heart disease based on the age, gender, pulse rate etc. The proposed system is based on the Neural Networks which has better accuracy based on the given data to predict the heart disease. It is a reliable and handy system to predict the vulnerability of a heart disease based on the basic symptoms given to it.

Application of Machine Learning in Disease Prediction:

The applications of Machine Learning have been increasing gradually in the medical field. The data contributed is used in the improvement of the classification and recognition systems and the data is used to train and test the model by which the increase in the survival rates has been seen. The classification algorithm can be done based on some diseases and the symptoms to get the required predicted output as the most accurate disease with which those symptoms are related to. There are different databases that contribute to the model to apply the classification based on them. These results have strengthened the idea of using the machine learning in disease prediction.

3. METHODOLOGY

Most of the past systems use the combination of algorithms and some use the complex algorithms to achieve a better accuracy but those systems are computationally expensive and require more storage to compute and process the data and to predict the desired result. Some systems use the CNN, KNN, Random Forest, SVM, combination of algorithms and also there has been used multiple combinations of algorithms to improve the correctness of the algorithm. There are also some systems designed based on the Naïve Bayes algorithm such that it is up to some diseases only but not for more. The results may not be accurate for previous systems because the model may not be trained well. Because of this problem, the accuracy of a model will be low than it has to be. So the model needs to be trained on different kinds of data as a dataset can have multiple kinds of data and the training must include all kinds of data in it to give better results. Hence the training also needs to be done in a different manner.

Disadvantages:

1. Unfortunately, the accuracy of many systems is very low and some have achieved a better accuracy but those are computationally expensive and cannot handle huge data.

2. Some methods based on better algorithms but those needs to be improved for more diseases as they are limited to few diseases.

3. Some models have better performance like SVM and Decision Tree but a small change in data may affect the whole model and leads to inaccurate results.

Proposed system and Advantages

The proposed system takes the data from the dataset to train and test the model and then the model will be ready to process. The data processing will be done in multiple ways such that the data will be distributed evenly as there will be multiple kinds of data and while separating the data into training and testing parts, the training part may get same kind of data and the testing part may get the data that is not present in the training data hence the accuracy

of the model will be very low as compared to the actual accuracy that needs to be get using that model. To face this challenge, we use the StratifiedKfold to separate the data into multiple combinations and perform train and test and then we use that data to achieve the maximum accuracy of the model. The system uses Naïve Bayes algorithm to predict the disease and the accuracy is also more for the Naïve Bayes model as it is one of the best classification models and also it is computationally inexpensive as compared to other algorithms. We have also included other algorithms to check whether the Naïve Bayes gives the best result or not.

- The training and the testing of the model are done in a sophisticated manner such that the data will be divided into best possible combination so that the model will have a better accuracy while training and testing. The system also has the interactive webpage in which the patient can enter the symptoms to predict the disease.
- The data will be taken from the frontend and the data will be checked in the database for a better result if it is already processed by another user which reduces the burden for the system to recalculate the result by running the algorithm every time a user enters the symptoms to predict the disease even the symptoms are same which may increase the load on the system.
- The data will be taken by the backend in which further will be run by the machine learning algorithm in which the dataset will be used to run the training and testing part in the model and then the model will be given the symptoms as the input and the desired result will be taken as the output from the model. The inputs will be run after the data has been separated into better combination as the model will work better when it has been trained and tested thoroughly. Finally, the result will be displayed in webpage.

Taking the symptoms from the user and processing them in the backend server and then the data will be processes and then the model will be processed and then the output will be shown in the frontend page by using the HTML, CSS, JS. The web interface is used to take the inputs and show the output to the patient. The data will also be saved in the database so that to reduce the load to the system when same symptoms are searched again so that to avoid the repeated process with the same symptoms. When the same combination of symptoms are given, then the algorithm will not run instead the result will be returned from the database where the accurate result has already been predicted to reduce the load to the system. Thus, the system will be fast in predicting the result and it does not require any additional load balancers to optimise the system performance which makes the system computationally low in cost.

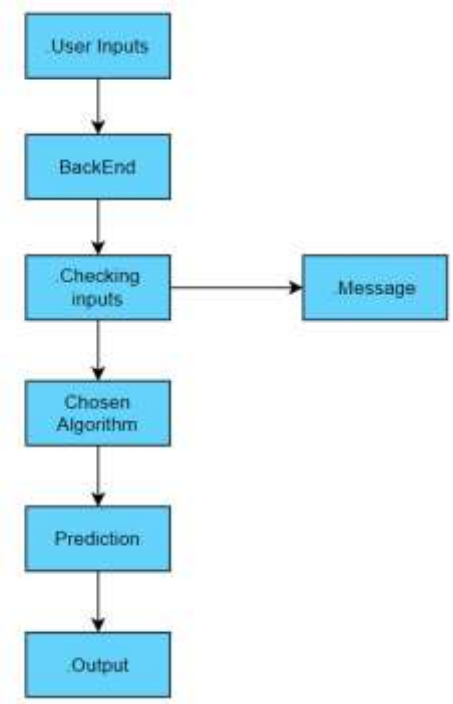

Fig.1: System architecture

MODULES:

We created the following modules for this project.

1. Enter the symptoms in Web Interface.

2. Take the symptoms from Webpage to Backend and process it to Machine Learning Algorithm.

3. Train and test the model and then give the inputs.

4. Prediction of the Run.

5. Show the result in the Web Interface.

4. IMPLEMENTATION

The Naïve Bayes implementation of this model is used to predict the disease based on the symptoms given by the user and the system consists of the Web-based user interface which is used by the user to interact with the system to get the required output for the user. The process is divided into different modules such that the interface will be separate and the backend will be separate.

In the system, the dataset will have the columns as symptoms and the rows will be of a specific disease with a value of either 1 or 0 which tells us that whether that symptom is related to that disease or not. The diseases will be replaced with the numbers to get the result in a numpy array. So, to get the result, we have to modify the dataframe in the model.

Also, the dataset may contain null values but mostly it will not affect the system because the values are either 0 or 1 so it will not be a problem as the null values can be taken as 0 so that the performance and the accuracy of the model can remain same and the prediction can be mostly accurate.

```
df.head()
```

	back_pain	constipation	abdominal_pain	diarrhoea	mild_fever	yellow_urine	yellowing_of_eyes	acute_liver_failure	fluid_overload	swelling_of_stomach	...
0	0	0	0	0	0	0	0	0	0	0	...
1	0	0	0	0	0	0	0	0	0	0	...
2	0	0	0	0	0	0	0	0	0	0	...
3	0	0	0	0	0	0	0	0	0	0	...
4	0	0	0	0	0	0	0	0	0	0	...

5 rows × 96 columns

Fig.2: First five rows of Dataset

Here it shows the data where the dataset contains the values of 0 and 1. This is only the first five rows hence there is no value of 1 as those were not the symptoms of the first five diseases.

```
In [6]: df.columns
Out[6]: Index(['back_pain', 'constipation', 'abdominal_pain', 'diarrhoea',
       'mild_fever', 'yellow_urine', 'yellowing_of_eyes',
       'acute_liver_failure', 'fluid_overload', 'swelling_of_stomach',
       'swelled_lymph_nodes', 'malaise', 'blurred_and_distorted_vision',
       'phlegm', 'throat_irritation', 'redness_of_eyes', 'sinus_pressure',
       'runny_nose', 'congestion', 'chest_pain', 'weakness_in_limbs',
       'fast_heart_rate', 'pain_during_bowel_movements', 'pain_in_anal_region',
       'bloody_stool', 'irritation_in_anus', 'neck_pain', 'dizziness',
       'cramps', 'bruising', 'obesity', 'swollen_legs',
       'swollen_blood_vessels', 'puffy_face_and_eyes', 'enlarged_thyroid',
       'brittle_nails', 'swollen_extremities', 'excessive_hunger',
       'extra_marital_contacts', 'drying_and_tingling_lips', 'slurred_speech',
       'knee_pain', 'hip_joint_pain', 'muscle_weakness', 'stiff_neck',
       'swelling_joints', 'movement_stiffness', 'spinning_movements',
       'loss_of_balance', 'unsteadiness', 'weakness_of_one_body_side',
       'loss_of_smell', 'bladder_discomfort', 'foul_smell_of_urine',
       'continuous_feel_of_urine', 'passage_of_gases', 'internal_itching',
       'toxic_look_(typhos)', 'depression', 'irritability', 'muscle_pain',
       'altered_sensorium', 'red_spots_over_body', 'belly_pain',
       'abnormal_menstruation', 'dischromic _patches', 'watering_from_eyes',
       'increased_appetite', 'polyuria', 'family_history', 'mucoid_sputum',
       'rusty_sputum', 'lack_of_concentration', 'visual_disturbances',
       'receiving_blood_transfusion', 'receiving_unsterile_injections', 'coma',
       'stomach_bleeding', 'distention_of_abdomen',
       'history_of_alcohol_consumption', 'fluid_overload.1', 'blood_in_sputum',
       'prominent_veins_on_calf', 'palpitations', 'painful_walking',
       'pus_filled_pimples', 'blackheads', 'scurring', 'skin_peeling',
       'silver_like_dusting', 'small_dents_in_nails', 'inflammatory_nails',
       'blister', 'red_sore_around_nose', 'yellow_crust_ooze', 'prognosis'],
      dtype='object')
```

Fig.3: Columns representing the symptoms

These are the complete set of symptoms that we are currently working on to predict the disease of the patient and these columns will have the value as 1 if it is associated with the disease present in that row.

Fig.4: Unique diseases and complete Dataset

We have taken up to 40 diseases to predict the disease based on the symptoms so that the dataset has been shown in the above figure.

Fig.5: Datatypes of each column

The datatype for each column is integer data in which we will store the value as 0 or 1 to get the association between a disease and a symptom. Also, all the names of diseases have been modified with integer values for the easier way of getting output. This implementation consists of an interface through which the input symptoms will be taken. The output will also be shown in the webpage. The inputs will be taken by the backend and the machine learning algorithm will run in the backend from the server.

The final output will be calculated by the server and the output will be given to the webpage with the use of backend and then the result will be shown in the frontend webpage.

The result will be calculated by the algorithm based on the most accurate trained model as the dataset will be separated into two parts i.e., training data and testing data. This system will calculate the best training and testing accuracy for different combinations of divisions and then take the most accurate data division for the training and testing. This model gives the most possible accurate results as the training of the model will be done in a better way than the previous systems which will improve the accuracy of the system and the Naïve Bayes algorithm makes the system computationally inexpensive.

5. EXPERIMENTAL RESULTS

Fig.6: Homepage of the website

The above figure shows the webpage of the system through which the user can give the symptoms to predict the disease based on the given symptoms. There are three different algorithms to run with and we can take the most accurate one to check the disease.

Fig.7: Data entered

Fig.8: Output of Naïve Bayes

Fig.9: Output of Random Forest

Fig.10: Output of Decision Tree

Fig.11: Insufficient symptoms

The above figures show the outputs of same symptoms for different algorithms and the last figure shows when the symptoms are not given up to three as giving less than three can make it difficult to predict the disease accurately.

6. CONCLUSION

The Disease Prediction using Naïve Bayes is a machine learning model used to predict the disease based on the symptoms given by the user. This is based on the Naïve Bayes algorithm which is one of the robust algorithms which gives a way better accurate results in classification using the symptoms given by the user and tells the disease the user has. There is a web-based UI provided to take the inputs and show the results. We have been showing the comparison to Naïve Bayes with Decision Tree and Random Forest as to show that Naïve Bayes has better accuracy than the other two algorithms. Also the algorithm will not run multiple times for the same set of symptoms but the stored result will be displayed based on the previous algorithm approach. The primary focus of this model is to keep the accuracy better even when the disease count increases.

REFERENCES

[1] Hamsagayathri, P., and S. Vigneshwaran. "Symptoms Based Disease Prediction Using Machine Learning Techniques." 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV). IEEE, 2021.

[2] Venkatesh, K., et al. "Identification of Disease Prediction Based on Symptoms Using Machine Learning." JAC: A Journal Of Composition Theory 14.6 (2021).

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

[4] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." IEEE access 7 (2019): 81542-81554

[5] Uddin, Shahadat, et al. "Comparing different supervised machine learning algorithms for disease prediction." BMC medical informatics and decision making 19.1 (2019): 1-16.

[6] Gavhane, Aditi, et al. "Prediction of heart disease using machine learning." 2018 second international conference on electronics, communication and aerospace technology (ICECA). IEEE, 2018.

[7] Hasija, Yasha, Nikhil Garg, and Soumya Sourav. "Automated detection of dermatological disorders through image-processing and machine learning." 2017 International Conference on Intelligent Sustainable Systems (ICISS). IEEE, 2017.

[8] Kohli, Pahulpreet Singh, and Shriya Arora. "Application of machine learning in disease prediction." 2018 4th International conference on computing communication and automation (ICCCA). IEEE, 2018.

[9] Katarya, Rahul, and Polipireddy Srinivas. "Predicting heart disease at early stages using machine learning: a survey." *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE, 2020.

[10] Patil, Mrunmayi, et al. "A proposed model for lifestyle disease prediction using support vector machine." *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2018.