International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-7, May 2020

Disease Prediction for the Deprived using Machine Learning

Krishna Kant Agrawal, Shivam Sharma, Shagun Tomar, Shubham Kumar

Abstract*: Our work aims for economical disease diagnostics, by asking the user for Prognosis and symptoms, accurate disease prediction has been strived for. In aspiration for social welfare, the cost of using the product built is almost free, the prediction can be done using any one of the six algorithms, five out of which are total free of cost for use, those five being KNN, Naïve Bayes, SVM , Logistic Regression, K Means Classifier. The one, that gives out predictions with most accuracy, i.e., Decision Trees Classifier, has been made paid, others are not to be paid for, for using.How this product would be functioning is simple: User logs in , openCV has been used for it, that brings the user to the section where user is briefed about models working on different algorithms, each algorithm having different accuracy, thus further, which model he/ she should choose. On choosing model of their choice, they fill their symptoms and prognosis, that yields them their final result of name of their disease.Services like these are greatly needed , looking at large many number of people in our society, who are unfortunately not able to afford them, when priced heavily, or even moderately. Such products can help save many a lives, notify sufferer about his chronic disease at early stage, inform about deficiency diseases, that are very controllable, if get known about, early.

Keywords* : Machine Learning, Classification, Logistic Regression, Decision Trees, K Means, K Nearest Neighbors, Support Vector Machines, Naïve Bayes.

I. INTRODUCTION

Deprived in our society are forced to live without even the most basic of medical facilities. Many a times their sickness starts, and remains at very controllable stage for quite long, where it can be dealt with, without even shelling out much heavy money. This, being the stage that is the most deterministic of patient's medical case, is very crucial, which can either be followed by his better condition, if he gets to know about his disease or, deteriorate, happening in case if he doesn't get to know about what he is suffering from. Economical diagnostics can help save many a life, if they, in form of a product are made available to those who need it.

Revised Manuscript Received on May 30, 2020.
***** Correspondence Author
Dr. K K Agrawal*, Department of CSE, ABES Institute of Technology, Ghaziabad, India. Email: kkagrawal@outlook.com
Shivam Sharma, Department of CSE, ABES Institute of Technology, Ghaziabad, India. Email: shvm1681998@gmail.com
Shagun Tomar, Department of CSE, ABES Institute of Technology, Ghaziabad, India. Email: risabhtg@gmail.com
Shubham Kumar, Department of CSE, ABES Institute of Technology, Ghaziabad, India. Email: shubham.80104.sk@gmail.com

Using Machine Learning, this service can be made available for social welfare, as an attempt to help those in need of a helping hand. A patient on getting infected by a disease, visits a doctor at least twice, before he is recommended for a diagnostic test, many don't even make those two visits, in reluctance of having to spend money on doctor's fees. For example, rising of body temperature, may be treated by the sufferer as common fever with flu, though it actually might be dengue, that, on the very day of suspicion of it being there, if it is tackled with, platelet count increasing fluids like coconut water, can help prevent their suffering. Failing to detect it in time may cause crashing platelet count, resulting death.

These sufferings of the poorer section, to such diseases can only be seen rising as of now, looking at the alarming rise in levels of pollution and other factors. 4.6 million is the number of people across world who loose their lives to just air pollution, 485,000 become prey to water pollution, the figures are scary. By 2050, the figures are expected to reach 6.6 million. Diseases caused by them are eating up the world, these diseases , making people reach their end in such manner can be stopped, treating disease at time of their early detection. Milking this as an "opportunity", the diagnostic industry in world, is expected to reach net worth of 19.35 billion dollars, by 2022 . Something that should have been a right, is being served as a "luxury". Machine Learning can help big time, in reducing this wrong going thing, it can shape out a product, that as a service can be used reduce the misery.

Predictive disease applications are widely being worked upon, they can help give way more economical diagnostic solutions, Machine Learning and Deep Learning are going to play a great role in it. Literature survey done showed that a lot of work has already been done in a specific disease restrained diagnostics, ie, if on suspicion of cancer being there, detection of what type of cancer is there. Diagnostics in wider manner, ie, diagnostics of general diseases , by taking as input, the general symptoms is something that hasn't been worked upon a lot. This concept on being properly worked upon, would yield a general physician doctor in form of an application. On being optimized, by including along, the set of general medicines prescribed, we can have the conceptualised general physician doctor in our cell phones.

II. LITERATURE REVIEW

[1] Heart disease, being one of the most common diseases , is attempted to, be detected, if user has any. The product built was an Android application, that took information from user in form of symptoms.

It gave messages to alert, in case if a disease was detected along with, information about doctor nearby. K Nearest Neighbors was used for classification of disease.

[2] A combined detection system for diseases related to diabetes, kidney and liver was made, using datasets for each. Algorithms applied in project were Support Vector Machines and Random Forests. The accuracy obtained was measured by precision, accuracy score, recall value.

[3] Heart problems, that have engulfed lives in such large numbers in the world, were predicted , using a dataset that had information of BP, smoking, hypertension, diabetes. Algorithms like KNN , Naïve Bayes , SVM and Decision Trees were used. Out of these algorithms, the one that showed highest accuracy, was taken to build the ultimate final working model.

[4] Cardio Vascular diseases were predicted, using Machine Learning algorithms like KNN, SVM, Decision Trees, Naïve Bayes

[5] K Means Classifier algorithm was used to detect the disease someone was suffering from. It had a very good accuracy score, of 0.95.

[6] A project that could solve absence of a doctor, using Machine Learning algorithms was built. It even provided a facility of interacting with a doctor, if he physically wasn't there. It is a very major problem in rural areas, that doctors being a part of govt. hospitals and clinics rarely stay there and serve, owing to lesser money making opportunity there. They mostly move to urban areas for making heavy sums of money. Left behind are those villagers, who don't have any medical expert to consult, if they fall sick. Such problem of theirs was solved by making a product, that could help them interact with a doctor, on one being there or in case of his absence, they could be given medical help like diagnosis, using Machine Learning algorithms.

[7.] Prepared Machine Learning model worked on K Nearest Neighbors algorithms, it worked on structured data of daily life habits like smoking, age, weight , gender and unstructured data had records of patients' narration of medical problem, doctor's interrogation and diagnosis. K Nearest Neighbors, and finally CNN were used in this project.

[8]. This work was aimed to detect heart ailments using classification algorithms like Decision Trees Classifier and Naïve Bayes algorithm, with dataset by Cleveland Heart Disease. Patterns were noted, as on what led to heart attack, with all the accuracies being monitored upon.

[9.] This research paper focused on predicting whether breast cancer, one was suffering from, was malignant or benign. Benign ones are relatively easier to treat, as cells in it do not propagate this disorder to neighboring ones. It used two datasets, one, Wisconsin Breast Cancer Dataset and other by UCI. It used 7 algorithms- Naïve Bayes, Decision Trees, SVM, MLP, K Nearest Neighbors, Random Forests and Bayes Net. Weka software was used for this prediction. For analysis of results, accuracy score, Precision value, F score, Recall value and ROC Area were used. Also, True Positives, True Negatives, False Positives, False Negatives helped determine the best performing algorithm.

[10.] This research paper focused on detecting presence of unbalanced quantity of High Density Lipoprotein and Low Density Lipoprotein (good cholesterol and bad cholesterol respectively). How bad cholesterol blocks the arteries, making sufferer prone to heart attack, was well explained. The dataset used for making predictions was Pima- Indian dataset, having 76 attributes, 10 out of which were, selected during data preprocessing, to train the models. There were four algorithms used for prediction purpose, namely Simple Multilayer Perceptron, Simple Naïve Bayes, Decision Trees, J48. These algorithms showed accuracies of 91%, 77%, 85% and 84% respectively.

[11.] This research paper predicted cardiovascular diseases by using datasets from laboratories of Jakarta, Indonesia, huge dataset of 60,589 records was used, with 38 medical attributes. In order to be able to get to these results, tests of blood, cholesterol etc were used. For prediction purpose, attributes such as age, urea, gender ,creatin, uric acid, cholesterol, triglyceride, HDL, were taken under consideration. For making these predictions, Naïve Bayes algorithm was used..

III. WORKING METHODOLOGY

Step1:Patient is supposed to log in the system. OpenCV has been used for purpose of login by face recognition.

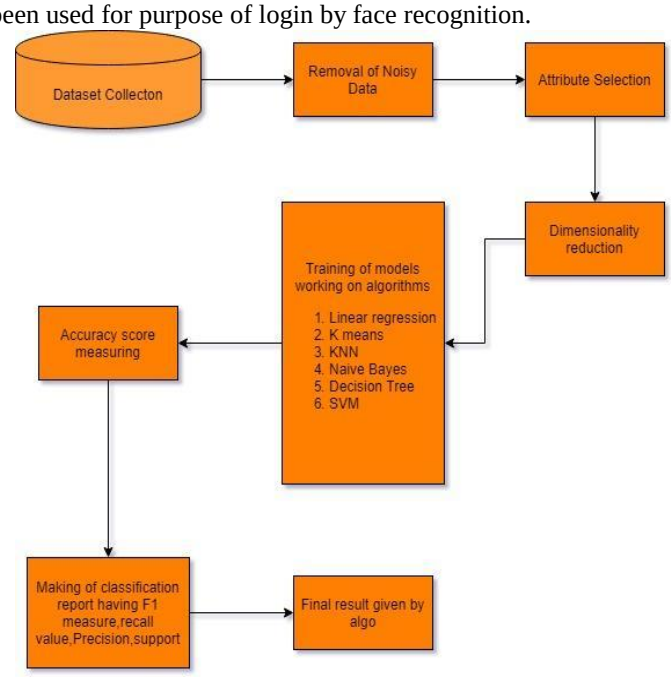

Fig.1. Working Methodology

Step2:User is then briefed about accuracies of models for prediction. He then can choose the model he wants to go by. Out of six algorithms, five have been kept total free for use, those five being Naïve Bayes, K Means,

K Nearest Neighbors, SVM, Random Forest .Model built on Decision Trees has been kept chargeable, with very minimal fees.

Step3: User then opts for the model, working on algorithm, he wants to go by.

Step4: User then has to feed in the prognosis, symptoms as inputs, in order to get to know which disease he is suffering from.

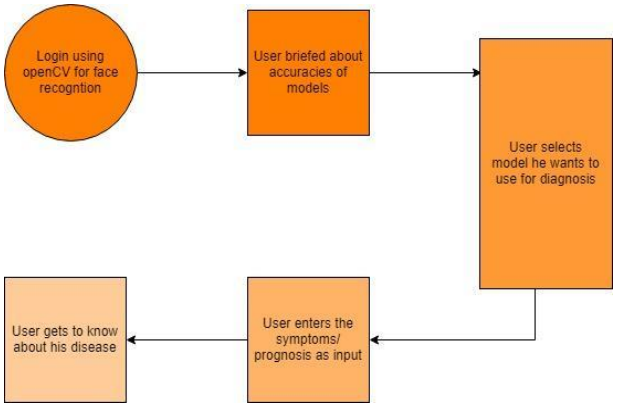

Fig. 2.Work Flow Diagram

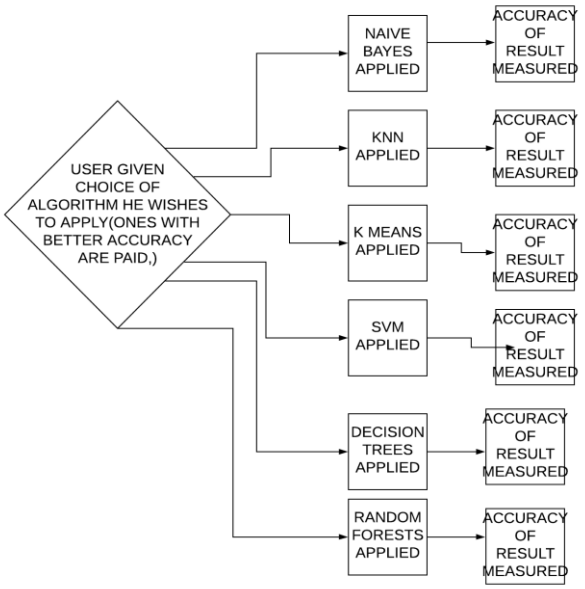

Fig. 3.Working of choices on algorithm

A.Algorithms Applied:

1) Support Vector Machines:

Support Vector Machines is a technique in machine learning that falls under category of supervised learning. It draws a decision boundary for classifying data points into classes. This decision boundary can be linear as well as non linear.Linear decision boundary helps better, when the data points are lying in a clean way, with their respective belonging classes. Non linear decision boundaries are used for data points that are lying in rather a much, messier manner. Along this decision boundary, run parallelly ,two lines on either sides of it. These two lines pass very closely with the most outward lying points of either classes, these lines are closest to these points, these points are called support vectors.

2) LOGISTIC REGRESSION

It is a regression technique, but can even be made to use in classification. It has a nonlinear relationship of dependent variables with the independent variables. Its values of dependent variable can be in the form of 0's and 1's.

3) NAÏVE BAYES

Naïve Bayes algorithm works on Baye's theorem.It has a bunch of algorithms but all work on the principle that every feature being analyzed is independent of each other. It treats every feature as totally independent, with every feature being of equal importance. It uses Baye's formula:

$$P(A/B) = P(B/A) * P(A)/P(B) \quad (1)$$

where A and B are two events.

4) Decision Trees

It is a very powerful algorithm; its functioning is in form of a tree. It is has branches in form of decision, its nodes are labels of test and branch represents outcome of the test. They can handle large attribute of data. The attributes from dataset need to be divided in recurrence further to have all the attributes and sub attributes.

5) K Means

K Means clustering is an unsupervised machine learning algorithm. It starts clustering with a number of centroids, fixed. Initial most points being the initial most clusters, their following points are calculated for their Euclidean distances, the following points are put in cluster of that centroid, whose Euclidean distance happens to be lesser. The newest addition to the cluster, the latest point and the current centroid are then calculated arithematic mean of, that yields the latest centroid. Following points are then clustered accordingly, in accordance with their Euclidean distance. The process then carries on until all points are clustered.

$$\text{Euclidean distance} : \sqrt{((x - x')^2 + (y - y')^2)} \quad (2)$$

where x is value of a data point
x' is value of centroid
y is value of a data point
and y' is value of centroid

6) K Nearest *Neighbors*

K Nearest neighbors algorithm is a supervised learning algorithm. It can be used for classification as well as regression. It uses a basic idea of forming a specific number of clusters, without any basic assumption of distribution of data.

Classification in it happens with reference to what class the neighbors of an unclassified point lie in, it is made to fall in that class. Looking for the neighbor, that would be used for classification, happens by measuring the distance using formula:

$$\sqrt{((p1 - q1)^2 + (p2 - q2)^2)} \quad (3)$$

Where points happen to be (p1, q1) and (p2, q2)

Table-I: Classification report of algorithms

Algorithms	Precision	Recall	F1 measure
DT	94.59%	91.62%	95.45%
Logistic Regression	82.21%	80.43%	89.91%
Naive Bayes	80.5%	78%	79.20%
SVM	82.4%	85%	80%
KNN	87%	81.74%	85.57%
K Means	85.64%	82.51%	89.43%

B. Description Of Dataset

Dataset used was taken from Kaggle. It is a huge, with 4920 rows of prognosis and 133 columns of symptoms.

The disease names in analysis were ranging from skin diseases like fungal infection, to chronic ones like AIDS. Number of symptoms under analysis were also very huge, they included the ones, as non serious seeming ones as runny nose, to as serious ones as coma.

C. Performance Measures Used

1) Accuracy Score

It is the simplest way of measuring accuracy, it mathematically is, number of correct predictions divided by total number of predictions. Its value lies between 0 and 1.

$$\frac{(\text{true positive} + \text{true negative})}{(\text{true positive} + \text{true negative} + \text{false positive} + \text{false negative})} \quad (4)$$

2) Confusion Matrix

Confusion matrix is a convenient measure of performance of algorithm, it forms a matrix of predicted values that turned out to be true or false, along with actual values, that are true, or false.

3) Classification Report:

a) Precision

It is a measure of how well, positives of dataset are being predicted. It is, how much of all the predictions that were made as positive, actually turned out to be positive.
Mathematically:

$$\frac{(\text{true positives})}{(\text{true positives} + \text{false positives})} \quad (5)$$

b) Recall

It measures of how well those predictions, that were said to be positive , actually turned out to be.

$$\frac{(\text{true positive})}{(\text{true positive} + \text{false negative})} \quad (6)$$

c) F1 SCORE

F1 score uses precision and Recall value for its calculation, it is used to set the right balance, between true positives and true negatives while making calculations for performance. Its mathematical formula:

$$\frac{(2 * \text{precision} * \text{recall value})}{(\text{precision} + \text{recall value})} \quad (7)$$

IV. RESULT AND DISCUSSION

The result of observations of different models put to use shows decision trees to be the best performing algorithm, while Naive Bayes performed less accurately. That justifies the reason why it has been made paid, though with very minimal fees. Performance of all the algorithms is shown below

Table-II: Accuracy scores obtained

ALGORITHM	ACCURACY
DT	0.91
KNN	0.89
SVM	0.82
NAIVE BAYES	0.75
LOGISTIC REG	0.83
K MEANS	0.85

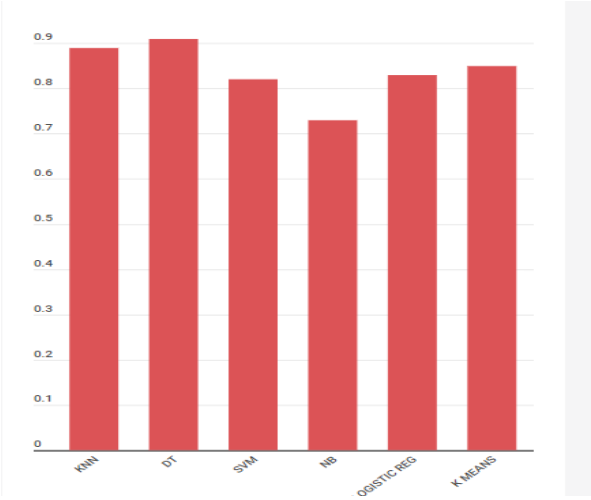

Fig. 3.Comparison of accuracies

V. CONCLUSION & FUTURE WORK

Our work can contribute greatly in healthcare and social work. Early disease detection is a major help in fight against it, saving money, time, and lives. Striving for excellence our future work would include:

1.To attach contact details of a specialist doctor, for treatment of the certain disease, that the user is detected with.

2. On detection of non-chronic diseases, for example, Fungal Infection, the user should have the flexibility of saving his/her money , by asking about treatment on chat, rather than, having to spend money and time for visiting clinic. Efficient chat feature would to be added along, ahead.

3. To develop a section in product, specifically for recommendations, do's and don'ts, in case user is detected with a mild disease like fungal infection.

REFERENCES

1. Heart Disease Prediction System pubished by Mrs. Jayshree LK et al. at IRJCS.

2. An algorithm for predictive data mining by Vivek Sharma et al at IJCSIT.
3. Analysis of Heart Disease Prediction Techniques by M. Marimuthu , S Deivarani.
4. Heart Disease Prediction using Machine Learning Techniques: A survey at IJET by V. V. Ramlingam et al.
5. Disease Prediction Using Machine Learning at IRJET by Akash C M Jagade et al.
6. Medicine Prescription Using Machine Learning by S Suma et al at GRD journals.
7. Prediction of probability of disease based on symptoms using Machine Learning Algorithms by Harini DK et al at IRJET.
8. Prediction of Heart Disease Using Machine Learning Algorithms by Sonam Nikhar et al at IJAEMS.
9. Comparative Study of Classification Techniques for Breast Cancer Diagnosis by Ajay Kumar et al.
10. Performance Evaluation of Classification Techniques in Diagnosing the Risk Factors for CVD by Dr. R Latha et al.
11. Detection of Cardiovascular Disease Risk's Level for Adults Using Naive Bayes Classifier by Eda Mirinda et al.

AUTHORS PROFILE

Dr. Krishna Kant Agrawal completed his Ph.D. from IIIT Allahabad, currently working as Professor in Department of Computer Science & Engineering, ABES Institute of Technology, Ghaziabad, has more than 40 research publications. ACM Professional Member.

Shivam Sharma is pursuing Bachelors in Technology in Computer Science from ABES Institute of Technology, currently in fourth year. Has worked in field of Data Science Analytics and Machine Learning, He has pursued internships, undergone rigorous trainings and developed projects in this domain. A pretty good coder he is in Python.

Shagun Tomar is currently pursuing B.Tech in Computer Science from ABES Institute of Technology, currently in final year. He had done projects in Java ,Python, Django & Tkinter.

Shubham Kumar is pursuing Btech in Computer Science from ABESIT, He has gathered rich enough experience in development related works using frameworks like django. He is trained in Python and Java coding.