iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VI **Month of publication:** June 2020

DOI: http://doi.org/10.22214/ijraset.2020.6183

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

Multiple Disease Diagnosis using Two Layer Machine Learning Approach

Aniket Dere[1], Omkar Patil[2], Paras Sutar[3], Tanmay Shende[4], Mahesh S Shinde[5]
[1, 2, 3, 4,] *B.E Students,* [5] *Assistant Professor, M.E.S College of Engineering, Pune, Maharashtra, India*

Abstract: ***As everyone knows healthcare is a basic human need. In a country as big as India there is shortage of qualified doctors for serving the large population. Remote parts of the country are still not able to afford and access quality healthcare. A person may not be able to consult with specialist doctors. Second opinion regarding someone's symptom is important for accurate diagnosis. With advancement is technology and increase in access of internet it is imperative to augment the traditional approach towards healthcare. Machine learning can be used to train a system based on symptoms of previously diagnosed patients. This system will be free of any biases and diagnose solely based on factual data. There is a need of a remote diagnosis system. The focus of this project is to aid and help a medical professional to verify and diagnose the patient with certainty using the symptoms provide by them.***

Keywords: Healthcare, Machine Learning, Diagnosis, Ensemble, Prediction

I. INTRODUCTION

Healthcare is a basic human need in the modern age. Currently India has in total 14000000 doctors practicing in India. Access to healthcare is an important issue in any developing country like India. Access Can be defined as the ability to receive services of quality at marginal cost and convenience. Provision, utilization and attainment are the factors in which India is lacking. Provision that is supply leads to utilization causing attainment of service. However, there for now exists a huge gap between these factors, leading to a damaged system with insufficient access to healthcare. Differential distributions of services, power, and resources have caused difference in healthcare access. Access to hospitals depends on many factors such as gender, socioeconomic status, education, wealth, and location of residence. Furthermore, inequalities in financing healthcare and distance from healthcare facilities are impediments to access. Additionally, there is a lack of sufficient infrastructure in areas with high concentrations of poor demography. Large numbers of tribes that live in isolated and these areas often have low numbers of medical professionals. Finally, health services may have long wait times or consider ailments as not serious enough to treat. Those with the greatest need often do not have access to healthcare.

Health is always a priority even before technology exists. Healthcare domain provides a lot of scope for research as it has tremendously evolved. There is a necessity of upgrading the existing Healthcare technology by embracing digitization of medical information.

Machine learning is another emerging and trending approach which closely works to solve the real time problems. Currently in the health care domain various data mining methods are used to find interesting pattern of disease using statistical medical data with the help of machine learning algorithms. Machine learning approach can be applied for prediction of diseases and provide automated diagnosis under the validation of professional doctor.

The individuals facing the problem of accessing healthcare at affordable prices. The basic aim of this project is to solve this problem with suitable accuracy and help the individual live a healthy life.

The basic idea of this project is building an application to predict multiple diseases. Providing an efficient and quality healthcare to a large population is a difficult task.

The expert system is trained using a generic dataset consisting of many diseases.[1]

First the symptoms of patients are given to the application. The ensemble technique diagnosis. After the first diagnosis a specific dataset of the disease is used to further make sure the results are accurate.

Ensemble technique is used with three supervised machine learning algorithms to provide high accuracy and to be closer to the prediction. In machine learning, boosting is an ensemble meta-algorithm for primarily reducing bias, and also variance in supervised learning, and a family of machine learning algorithms that convert weak learners into strong ones.

©IJRASET: All Rights are Reserved

II. RELATED WORK

A. *Chatbot for Disease Prediction and Treatment Recommendation using Machine Learning*

This paper proposes a chatbot for the prediction of diseases. The chatbot interacts with the user and gets the symptoms and health issues of the individual. The symptoms are then identified by the use natural language processing (NLP). These symptoms are then fed to KNN algorithm which predicts the disease.

Medical chatbot which can be used to replace the conventional method of disease diagnosis and its treatment. Chatbot can act as a doctor. The chatbot is a user application. The user of this application specifies their symptoms to the chatbot and in turn, chatbot will specify the health measures to be adapted. General information about symptom and diseases is available in the dataset and therefore the chatbot instance is able to provide information about the disease and treatment to the user. After analyzing the symptoms of the different users, it predicts the disease to the user and provides a link where details about the treatment are visible.[1]

B. *Efficient and Privacy Preserving Online Medical Prediagnosis Framework Using nonlinear SVM*

The authors propose an efficient and privacy preserving online medical pre diagnosis framework called as eDiag. It uses non-linear kernel support vector machine. eDiag is an online framework which accepts encrypted queries that are performed directly at service provider without decrypting the query. With eDiag, the sensitive personal health information can be processed without privacy disclosure during online prediagnosis service. Specifically, based on an improved expression for the nonlinear SVM, an efficient and privacy-preserving classification scheme is introduced with lightweight multiparty random masking and polynomial aggregation techniques. The encrypted user query is directly operated at the service provider without decryption, and the diagnosis result can only be decrypted by user.[2]

C. *Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method*

This paper presents a majority voting ensemble method that is able to predict the possible presence of heart disease in humans. The prediction is based on simple affordable medical tests conducted in any local clinic. The data was split into training and testing sets into a ratio of 80% training data and 20% testing data.

Finally, an ensemble classifier is applied where the classification is done based on the majority vote of the models (hard voting.) The voting occurs when each model makes a prediction for each instance and the output prediction is the one that receives more than half of the votes.[3]

D. *Diagnosis of Alzheimer's Disease using Machine Learning*

This paper proposes making use of machine learning algorithms to process the data obtained by neuroimaging technologies for detection of Alzheimer's in its early stage. Advances in medical technologies have given access to better data for confirming symptoms of various diseases in early stages. Alzheimer's disease is chronic condition that leads to degeneration of brain cells leading at memory loss. CT, MRI, PET, EEG, and other neuroimaging techniques are suggested for patients with cognitive mental problems such as confusion and forgetfulness, also other symptoms including behavioral and psychological problems. Clustering algorithms have also been implemented along with Fuzzy interference system. The algorithms Logistic Regression, Support Vector Machine, Gradient boosting, Neural Network, K-Nearest Neighbor, Random Forest are implemented.[4]

E. *Computer-Aided Diagnosis System for Rheumatoid Arthritis using Machine Learning*

This paper proposes the finger joint detection method and the mTS score estimation method using support vector machine. Rheumatoid Arthritis' progress can be evaluated by widely used measure called Modified Test Score(mTS). The mTS score assessments on hand or foot X-ray image is calculated several times a year, and it is time consuming. This gives rise for the need of an automatic mTS score calculation system. This study expresses the rough shape of a finger joint using histogram of oriented gradients (HOG). The support vector machine (SVM) using HOG detects finger joints on the X-ray image, and the modified Total Sharp (mTS) score is estimated by the help of support vector regression (SVR).[5]

F. *A Modified SVM Classifier Based on RS in Medical Disease Prediction*

This paper proposes a modified Support Vector Machine(SVM) classifier based on RS for medical disease prediction. RS provides new scientific logic and research method for information and cognitive science, and also develops effective preprocessing techniques for intelligent information process. Relevant features influencing the medical disease are extracted. These features are used as the input vectors of SVM. The medical disease prediction model is conducted, which makes great use of the advantages of

©IJRASET: All Rights are Reserved

RS in eliminating redundant information and take full advantage of SVM to train and test the data. RS, as an anterior preprocessor of SVM, can find out these relevant features influencing the medical disease. By comparing with other machine learning algorithms, it can be implied that the training rapidity and accuracy of the proposed model are both evidently modified in medical disease prediction.[6]

G. Disease Prediction using Hybrid K-means and Support Vector Machine

This paper proposes a disease prediction method using hybrid K-means and Support Vector Machine algorithm. The hybrid K-means algorithm is used for reducing the dimensionality of the given dataset. This reduced dataset is then fed to Support Vector Machine as input. The simulation is performed on MATLAB.[7]

III. PROPOSED ARCHITECTURE

A. Input

The input for training the model is in the form of dataset. A dataset with multiple diseases is used as a preliminary dataset. Dataset for individual diseases like diabetes, thyroid, cardiovascular, hepatitis, liver disorder and chronic kidney are used further to train individual models. Input symptoms are gathered from users through a web application.

Dataset	Attributes	Tuples
Multiple	107	1398
Cardiovascular	12	3422
Chronic kidney	25	401
Diabetes	9	769
Hepatitis	20	156
Liver disorder	11	584
Thyroid	25	1003

B. Data Preprocessing

The data in the datasets is raw data. The gathering of this data is loosely controlled due to which it may contain some discrepancies. Data preprocessing is used to get a data without missing values, out of range values, impossible data combinations, null values etc. The preprocessing of data is done using python's pandas library.

C. Classification Algorithm

Data classification takes place in two steps. First, the classification model is trained using data then the model is used to classify for a certain input data. The three algorithms used are Decision tree, Naïve Bayes and Random Forest.

1) *Decision Tree:* A decision tree is a flowchart-like tree structure, in which each internal node (non leaf node) denotes a test on an attribute, a branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node. Given a tuple, X, for which the associated class label is unknown, the attribute values of the tuple are tested against the decision tree. A path is traced from the root to a leaf node, which holds the class prediction for that tuple. Decision trees can easily be converted to classification rules. [10]
2) *Naïve Bayes:* In machine learning, naïve Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Baye's theorem with strong (naïve) independence assumptions between the features. They are among the simplest Bayesian networks models. But they could be coupled with Kernel Density estimation and achieve higher accuracy levels. Naïve Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum likelihood training can be done by evaluating a closed form expression which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.[8][9]

Bayes theorem is stated as

$P(c|x) = P(x|c)*P(c)/P(x)$

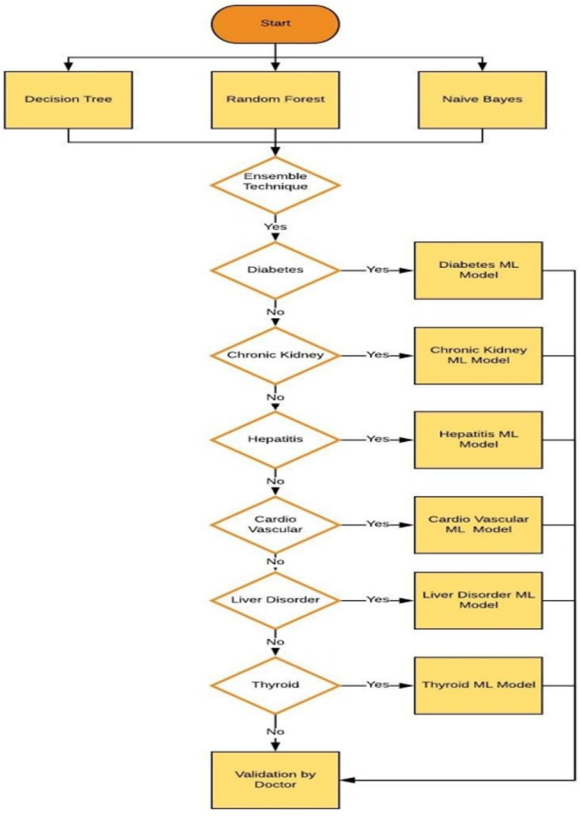
Where,

$P(c|x)$ is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).
$P(c)$ is the prior probability of *class*.
$P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
$P(x)$ is the prior probability of *predictor*.

©IJRASET: All Rights are Reserved

3) *Random Forest:* A random forest is a set of decision trees built on random samples with a different method for splitting a node: Instead of looking for the best choice, in such a model, a random subset of features (for each tree) is used, trying to find the threshold that best separates the data. As a result, there will be many trees trained in a weaker way and each of them will produce a different prediction. There are two ways to interpret these results; the more common approach is based on a majority vote (the most voted class will be considered correct). However, scikit-learn implements an algorithm based on averaging the results, which yields very accurate predictions. Even if they are theoretically different, the probabilistic average of a trained random forest cannot be very different from the majority of predictions (otherwise, there should be different stable points) therefore the two methods often drive to comparable results.[12]

4) *Ensemble Learning:* A particular computational intelligence problem is solved by strategically generating and combining multiple models, such as classifiers or expects. This is called Ensemble Learning. Performance of a model (classification, prediction, function approximation, etc.) is improve or the likelihood of an unfortunate selection of a poor one is reduced by the use of Ensemble Learning. Combination of multiple models constitute as Ensemble Learning. Bagging, which stands for *bootstrap aggregating*, is one of the earliest, most intuitive and perhaps the simplest ensemble-based algorithms. Bootstrapped replicas of the training data are used to obtain diversity of classifiers in bagging. That is, different training data subsets are randomly drawn – with replacement – from the whole training dataset. A different classifier of the same type is trained using each training data subset. Individual classifiers are further combined by taking a simple majority vote of their particular decisions. For any given instance, the class which is chosen by the greatest number of classifiers is the ensemble decision. Subset of the training data for training each classifier or using relatively weak classifiers can be employed as additional measures as the training datasets may overlap. Similar to bagging, boosting also creates an ensemble of classifiers by resampling the data, which is then combined by process of majority voting. In boosting, resampling is strategically geared such as it provides the most informative training data for each consecutive classifier. In essence, each iteration of boosting creates three weak classifiers: the first classifier $C1$ is trained with a random subset of the available training data. The training data subset for the second classifier $C2$ is chosen as the most informative subset, given $C1$. Specifically, $C2$ is trained on a training data only half of which is correctly classified by $C1$, and the other half is misclassified. The third classifier $C3$ is trained with instances on which $C1$ and $C2$ disagree. The three classifiers are combined by a three-way majority vote. [11]



©IJRASET: All Rights are Reserved

D. Data Storage

User's data is stored in MySql database. MySql is an open source Relational Database Management System (RDBMS). For handling of database operations Object Relational Mapping (ORM) framework is used. Object relational mapping (ORM, O/RM and O/R mapping tool) in computer science is a programming technique for concerting data between incompatible type systems using object-oriented programming language. This creates, in effect, a virtual object database that can be used from within programming language. ORM provides security from attack such as SQL injections.[13]

IV. RESULT

Classification Algorithm	Accuracy
Decision Tree	99.57 %
Naïve Bayes	98.92 %
Random Forest	97.49 %
Ensemble Technique	98.92 %

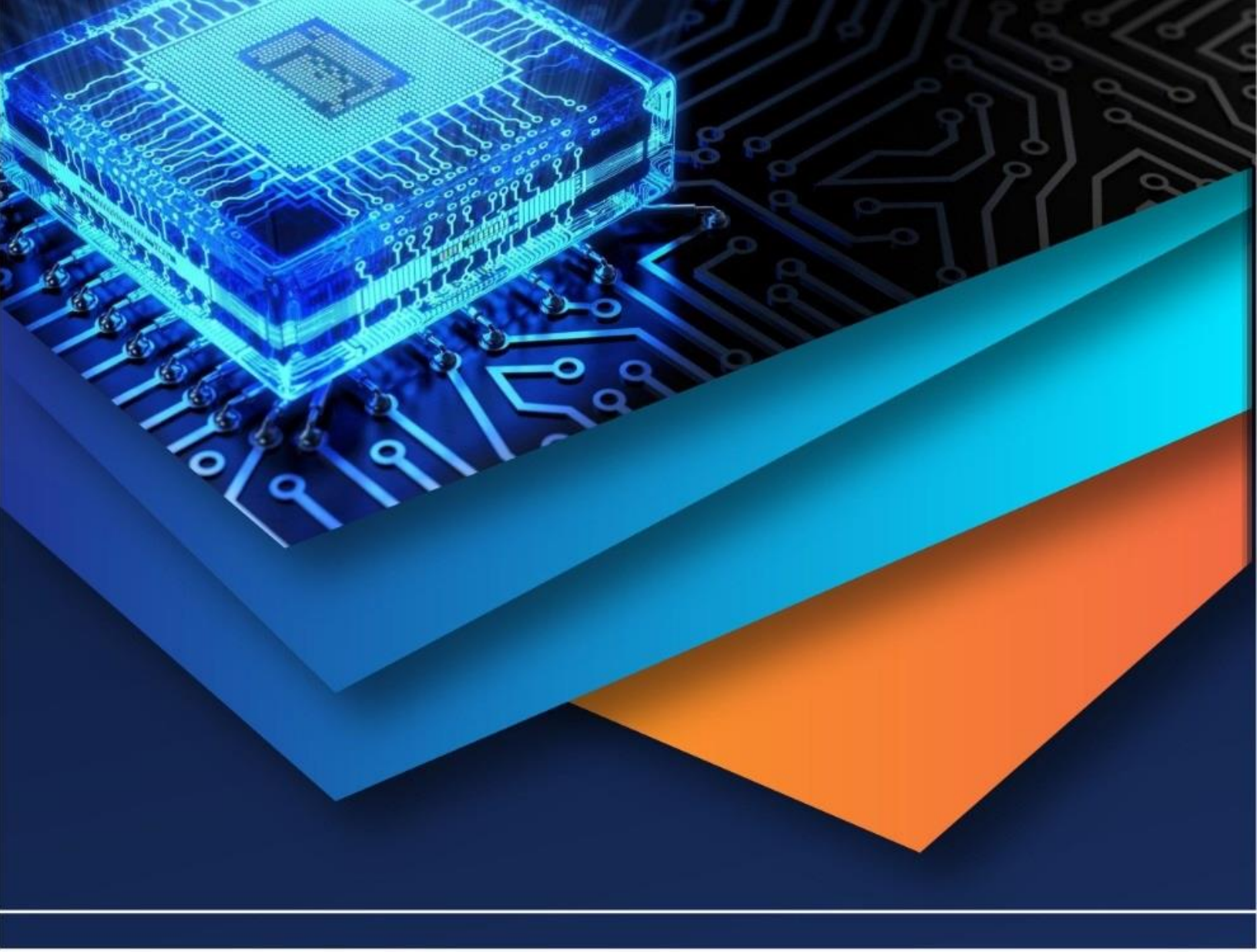
V. CONCLUSION

Here we proposed a healthcare system web application which uses three machine learning classification algorithms to predict the disease based on the symptoms and further classification of particular disease based on medical test. Boosting ensemble technique is used to boost accuracy of prediction. Application depicts the real world technique of detecting disease starting from symptoms to results of medical tests. In future the system can be improved by adding scrutiny of report functionality so that the user does not need to enter every value manually.

REFERENCES

[1] Rohit Binu Mathew, Sandra Varghese, Sera Elsa Joy, Swanthana Susan Alex, "Chatbot for Disease Prediction and Treatment Recommendation using Machine Learning", Third International Conference on Trends in Electronics and Informatics (ICOEI 2019) IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8
[2] Hui Zhu, Xiaoxia Liu, and Hui Li, "Efficient and Privacy-Preserving Online Medical Prediagnosis Framework Using Nonlinear SVM ", IEEE Journal of Biomedical and Health Informatics, VOL. 21, NO. 3, MAY 2017
[3] Rahma Atallah, Amjed Al-Mousa,"Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method ", IEEE 978-1-7281-2882-5
[4] Priyanka Lodha , Ajay Talele , Kishori Degaonkar, "Diagnosis of Alzheimer's Disease using Machine Learning ", 2018 Fourth International conference on Computer Communication Control and Automation (ICCUBEA)
[5] Kento Morita, Atsuki Tashita, Manabu Nii, Syoji Kobashi, "Computer-Aided Diagnosis System for Rheumatoid Arthritis using Machine Learning", International Conference on Machine Learning and Cybemetics,Ningbo,China,July 2017
[6] Guojun Zhang,"A Modified SVM Classifier Based on RS in Medical Disease Prediction ",2009 Second International Symposium on Computational Intelligence and Design
[7] Sandeep Kaur , Dr. Sheetal Kalra, "Disease Prediction using Hybrid K-means and Support Vector Machine ", IEEE 978-1-4673-6984-8
[8] Naïve Bayes available at : https://en.wikipedia.org/wiki/Naive_Bayes_classifier
[9] Naïve Bayes available at: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/
[10] Jiawei Han. Micheline Kamber. Jian Pei, "Data Mining : Concepts and Tecchniques", 3rd edition, ISBN 978-0-12-381479-1
[11] Bagging and Boosting available at: http://www.scholarpedia.org/article/Ensemble_learning#:~:text=Ensemble%20learning%20is%20the%20process,%2C%20function%20approximation%2C%20etc.)
[12] Giuseppe Bonaccorso , " Machine Learning Algorithms ", 1st edition ,ISBN 978-1-78588-962-2
[13] ORM available at : https://en.wikipedia.org/wiki/Object-relational_mapping

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)