International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Disease Prediction and Drug Recommendation Android Application using Data Mining (Virtual Doctor)

Vivek Mudaliar, P.Savaridaasan, Sachin Garg

Abstract: ***Data Mining is a method that requires analyzing and exploring large blocks of data to glean meaningful trends and patterns. In today's period, every person on earth relies on allopathic treatments and medicines. Data mining techniques can be applied to medical databases that have a vast scope of opportunity for textual as well as visual data. In medical services, there are myriad obscure data that needs to be scrutinized and data mining is the key to gain useful knowledge from these data. This paper provides an application programming interface to recommend drugs to users suffering from a particular disease which would also be diagnosed by the framework through analyzing the user's symptoms by the means of machine learning algorithms. We utilize some insightful information here related to mining procedure to figure out most precise sickness that can be related with symptoms. The patient can without much of a stretch recognize the diseases. The patients can undoubtedly recognize the disease by simply ascribing their issues and the application interface produces what malady the user might be tainted with. The framework will demonstrate complaisant in critical situations where the patient can't achieve a doctor's facility or when there are situations, when professional are accessible in the territory. Predictive analysis would be performed on the disease that would result in recommending drugs to the user by taking into account various features in the database. The experimental results can also be used in further research work and for Healthcare tools.***

Keywords: Data mining techniques, E-healthcare, locally frequent patterns, medical data mining, Symptoms, drugs

I. INTRODUCTION

The E-Healthcare framework is an end client bolster and this system enables clients to get indication on their disease and also drugs related to that disease. According to the World Health Organization ("WHO"), e-Health signifies "the utilization of data and correspondence advancements ("ICT") for well being". The definition, however exceptionally succinct, isn't extremely useful.

Manuscript published on 30 September 2019
* Correspondence Author
Vivek Mudaliar*, Information Technology Department, SRM Institute of Science and Technology, Kattankulathur,Chennai, Tamil Nadu
mail:vivekmudaliar81@gmail.com
P.Savaridaasan, Information Technology Department SRM Institute of Science and Technology, Kattankulathur,Chennai, Tamil Nadu
E-mail:savaridp@srmist.edu.in
Sachin Garg, Information Technology Department SRM Institute of Science and Technology, Kattankulathur,Chennai, Tamil Nadu
E-mail:Sgarg3497@gmail.com

The European Commission has advanced a more intricate meaning of E-Health. E-Health alludes to "instruments and administrations utilizing data and correspondence innovations that can enhance anticipation, conclusion, treatment, checking and administration". In this manner, the articulation e-Health might be securely said to incorporate the two apparatuses and administrations that utilization ICTs for purposes associated with well being.

Human services are considered as one the most recognized components of economies of the created nations where subjects request access to the high caliber and proficient care which is considered as one of the significant obligations of their administration. Social insurance is straightforwardly in charge of individuals well being and prosperity, it should be persistently created and change for better to bring not just effectiveness and efficiency for itself and abatement costs yet additionally to convey decent quality support of enhancing quiet security and nature of care. IT has been proposed as a basic apparatus in taking care of these issues and advancing better medicinal services. The framework includes making sufficient data accessible to specialists, medical attendants, and patients. The System enables the client to share their side effects and issues and given that investigates the malady and suggests close-by healing facilities, specialists and pharmaceutical. The System causes specialists to enter and see the quiet history and also other patient subtle elements.

The objective of Machine Learning in this paper is to through a Mobile Application framework adjust and gain from their encounter. Machine Learning approach incorporates the Mobile Application based framework into the medicinal services field keeping the end goal in mind to acquire the best and exact outcomes for the framework. Here the framework manages programmed recognizable proof of enlightening sentences from medicinal distributed by restorative diaries. Our primary point is to coordinate machine learning in the therapeutic field and assemble an application that can do naturally distinguishing and dispersing ailment and treatment-related data, encourage it additionally distinguishes semantic relations that exist amongst illnesses and medicines.

In the proposed work client will enter the various symptoms or side effects in the user interface of the application where the data will be processed by a continuous delivery system with the aid of the azure pipeline implementation. The side effects entered by the client are sent to the cloud through JSON (JavaScript Object Notation)parser where it would be arranged utilizing the Algorithm to make the further procedure simpler to locate the semantic catchphrase which recognizes the disease effortlessly and rapidly.

At that point, the semantic catchphrase found is coordinated with the put away therapeutic info database to recognize the correct sickness identified with that catchphrase show. Once the ailment identified, it is sent to the restorative database to separate the articles relating to that sickness. Our database contains rows with a one to one mapping from symptom to disease. To change this we create instances of all possible combinations of the symptoms for a disease. So that our machine learning model can learn the combinations of symptoms that lead to what disease. In our usage of our proposed framework, we have used various algorithms but the Apriori calculation gave us the most pre-eminent outcome. The issue explanation of the current framework was, it didn't distinguish the best ailment treatment but gave the most probable ailment. So the proposed arrangement utilized information mining ideas utilizing voting calculation to determine the issue furthermore, recommend the drugs with the highest average ratings through predictive analysis by the framework.

II. METHODOLGY

The main objective of this project is to develop an android application which uses some data mining algorithms, languages used will be python and java. The system takes the symptoms from the users which they are feeling at that moment and runs a Machine Learning algorithm in the cloud to detect the disease from which the user may be suffering. The System collects raw data from the user or consumer. As the massive amount of information is already available from healthcare websites, patients can easily compare the diagnosis done by their doctors and the related information which is already present on the internet. Also by accessing online support group chat system patients can communicate with other patients who are suffering from similar kinds of diseases, this way they can exchange information who might have suffered the same kind of symptoms. The system uses the provided data from the user and matches the symptoms already stored in the database. The database uses various data mining techniques and an intelligent algorithm.

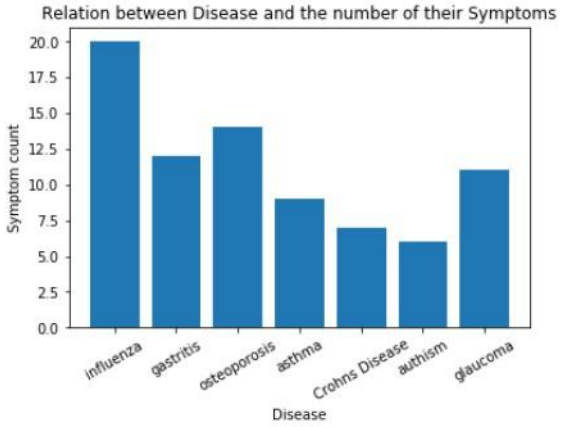
III. DATA GATHERING

The raw data is gathered from websites like mayoclinic.org, dataworld.org, and kaggle.com. The raw dataset which is a CSV file contains two columns the first one is disease and the other one is the related symptoms for that particular disease. Every disease contains at least 4-5 symptoms for the same and the data is then sent for the pre-processing so that the python code can be implemented efficiently.

Table [1]: Data Collected

Disease	Symptoms
Influenza	Fever, Chill, Headache, Sneeze
Asthma	Wheezing, Cough, Pleuritic pain
Gastritis	Vomiting, Abdominal pain, Intoxication
HIV	Fever, Cough, Diarrhea

The disease symptom dataset contains 2525 rows contains multiple diseases that counts upto 52 unique disease and for the drug dataset each disease has multiple drug which differ in rating and user count.



IV. DATA PRE-PROCESSING

After gathering the data in raw form the data is transformed into another .csv file which has indexing of every symptom mapped with every disease in 0 and 1's by the method of one-hot encoding. The diseases are made as rows and all the symptoms are made as columns like the table made below, the presence of every symptom for a particular disease is marked as 1 and its absence is marked as 0 all according to the dataset which we collected earlier i.e raw dataset.
From table [2] we can infer that the symptoms like fever, chill are present in Influenza that's why it is marked 1 and since Cough and Nausea are not present so it is marked as 0 according to the dataset which we collected same goes with Asthma and Diabetes as well.

Table [2]: Cleaned Data

Disease	Fever	Chill	Nausea	Vomiting
Influenza	1	1	0	0
Asthma	0	0	1	0
Diabetes	0	0	0	1

V. ALGORITHM IMPLEMENTATION

Data mining combines fact-based examination, machine learning, and database design to remove shrouded connections from vast databases.It uses two methodologies: Supervised learning and Unsupervised learning. In supervised, a training set is used to display parameters and in unsupervised learning no training set is utilized. In data mining each technique serves a different purpose depending on the objective of the modeling. The two most common modeling objectives are Classification and Prediction. We used a bunch of different algorithms for the diagnosis of the disease like the Decision tree, Naïve Bayes, KNN, Logistic Regression, and Gradient Boosting.

For these classification models we created all combinations of disease and symptom as we have one to one mapping of the disease and symptom, this process is not requisite in Apriori as it does that during its implementation. Due to this, we conclude that the association rules like APRIORI, CDA, etc. work more efficiently

A. Decision Tree:

This Algorithm is a choice help apparatus that uses a diagram which is tree-like and their conceivable results, which also includes possible occasion results, and utility. It is the type of approach which shows a calculation alone contains control articulations that are restrictive. We used the sklearn kit to implement a decision tree algorithm in python to predict the possible disease according to the symptoms given. We created the test, train data from the collected dataset and used accordingly. Our dataset has 2542 rows out of which we trained the 2000 rows and tested 542 rows remaining. The data is been read using variable_name=read.csv(“) and training is done by implementing the DecisionTreeClassifier() function and to fit the model using model.fit(). The accuracy achieved: 85-90%.

B. Naïve Bayes:

This algorithm forms the basis of different data mining and machine learning models. Naïve Bayes works on prediction modules to help predict the possible outcomes. We used python language for the implementation of it with the dataset we collected. Since it is present in package sklearn kit so importing and loading that is mandatory. Then the dataset is been read using read.csv() then 75% of the data is been trained and the remaining 25% is tested against it all of this done is using NaiveBayes() and predict() present in sklearn package and the predicted output is noted but the predicted data sometimes is correct and sometimes isn’t as it doesn’t know about the values which didn’t occur in that 75% of the data, the accuracy is more when the whole dataset is trained. Accuracy Achieved: 80-85%.

C. KNN:

K nearest neighbors is a prediction algorithm that works on the Euclidean distance concept. The distance is measured from different clusters the closer the distance the most likely the element is to belong to that cluster. We implemented this in python the dataset is been read using read.csv() and then the class library is loaded since KNN is present in it using library(class) and then the use of functions like model.train() which is used to train the dataset and the prediction model is made using KNeighborsClassifier() and the accuracy is seen, thus the prediction is checked. Accuracy achieved: 75-85%.

D. Gradient Boosting:

The gradient boosting algorithm can be explained by first introducing the algorithm named as AdaBoost. This algorithm begins by training a decision tree in which it assigns equal weight to each observation. After the first tree is evaluated, the weights are increased for those observations that are difficult to classify and lower the weights for the ones which are easy to classify .This was also implemented using sklearn kit which provides the function GradientBoostingClassifier() to implement on the train split of the data which was again 75%on train dataset and remaining 25% on the test dataset.Accuracy achieved: 90-95%.

E. Apriori:

This algorithm is a calculation for mining successive thing sets for Boolean rules. It is a "bottom up" technique, in which frequent subsets which works on one thing at any given moment. It is intended to work on database containing exchanges.

Key Concepts:

• Regular Occurring Item sets: Items that have minimum number of support which is denoted by Fi for the ith-item”.

• Apriori Characteristics: All possible regular occurring subset of item sets must be regular. The joining process includes finding Fj, set of fellow j-item sets can be obtained by Fj-1 joining with itself.

Steps Involved:

• Getting sets of frequent items:Items that are in a group that have the minimum support value then a condition which must be true “A subset of a regular item set should also be a regular item set”, i.e, if {YZ} belongs to a regular item set then the frequent item set should be both {Y} and {Z} – Repeatedly regular item sets need to found with cardinality which has range from 1 to the j-item set.

• Regular item sets should be used to form association metrics.

Pseudo Code:

• Joining: Candidate C k is formed by joining Fj-1with its own self.

• Cleaning: An item can be a subset of a frequent j-item set if only (j-1)-item set is frequent.

• Pseudo code part: Cj: Candidate set which is of size j Fj : Regular item set which is of size k F1 = {regular items}; for (j=1; Fj != Ø; j++) begin Cj+1 = candidates that is generated from F j; the increment of the count of all candidates for each transaction t in database in Cj+1 that are contained in t Fj+1 = candidates in Cj+1 with minimum support end return ∪ j L j. Accuracy achieved: Above 95%.

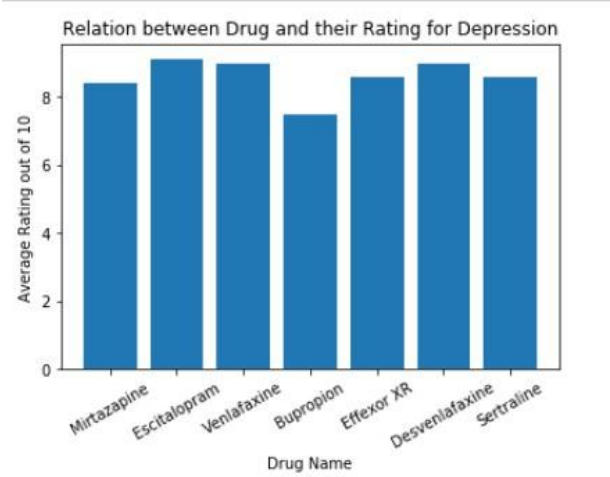
F. Predictive Analysis:

The disease prediction is done and the next step is to recommend drugs to the user. The disease symptom dataset and the drug dataset has been merged using pandas library in python. The drug dataset contains 5 columns which are disease, drug name, review, rating and useful count. We needed average rating of a particular drugs as there were 52 different diseases and each disease had multiple drugs. The average rating can be calculated using the weighted average method between the rating column and the useful count column.There is lot of similarity between the weighted arithmetic and ordinary arithmetic mean (average) except that in the weighted average some data points are more to contribute to the equation and in hte ordinary method each data points contribute equally for the final average.

Weighted Average Formula:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i},$$

W is the weight relative to a particular element x.



This is all done using pandas in python and a new dataset is created with the disease that are present in the first dataset and the average rating calculated from the drugs dataset along with the name of the drugs.This dataset will be included in the application and top 5 rated drugs will be recommend in a new activity in the android along with its reviews.

VI. ARCHITECTURE OF THE APPLICATION

This android application uses java language in software known as android studio.The application has two login pages one for user login(user interface for patient) and admin login(login for the admin).Here the machine learning code which is written in python would be uploaded to the cloud and the symptoms entered by the user in the application would be sent to the cloud in json format and the output will be retrieved in the same format and the most probable disease will be predicted.Also a full page information about the particular disease will be presented.The next activity page will contain the top 5 average rating drugs which are relevant to the disease.A feedback form will also be provided, so that if the patient checks with a real doctor then how accurate was the disease prediction. This would help improve the application in future.

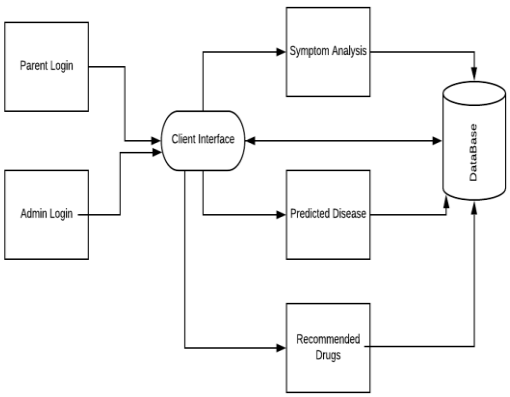


Fig 1: DataFlow Architecture

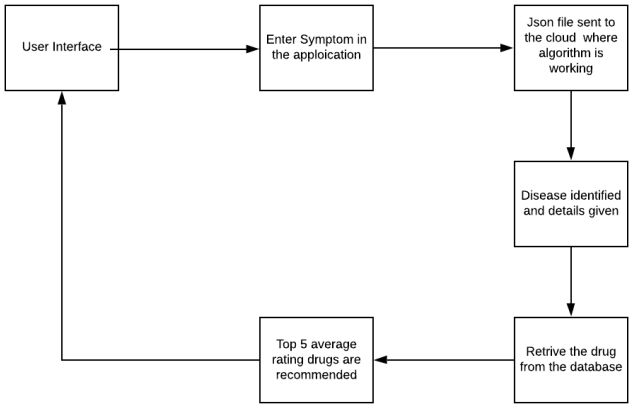


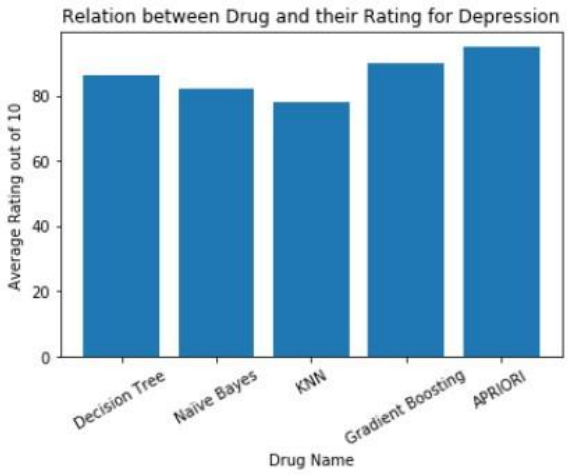
Fig 2:Proposed System

VII. RESULTS

In this paper we had proposed an android application which will make use of data mining for disease diagnosis. The system will prove useful in urgent cases where patient is unable to reach the doctor, for crisis cases that don't have specialists in a territory, amid crises that arise at late night and alsofor furthermore preparatory examination of the patients. This paper uses various machine learning algorithms and multiclass classification algorithms require to create subsets of every combination of disease and symptom.The Apriori algorithms creates subsets itself which is a more efficient way then creating the combinations manually using numpy and pandas library in python, so this algorithms gives the best accuracy.The drug prediction was done using predictive analysis between two columns of the the drug dataset which was merged with the first(disease symptom) dataset and using weighted average between the rating of the drug and useful count(number of people who rated that particular drug), the drugs with top ratings were recommended. This paper set up that while the current handy utilization of information mining in wellbeing related issues is restricted, there exists an incredible potential for information mining methods to enhance different parts of Clinical Predictions. Moreover, the unavoidable ascent of clinical information will expand the potential for information mining methods to enhance the quality and lessening the cost of human services.

VIII. DISCUSSION

The graph describes the comparison between the algorithms which and states the best algorithm among all the applied algorithms from which Apriori is the best option and it could be further improved if there are more data included in the dataset and less one to one mappings between the disease and the symptom. The accuracy could also be improved by taking feedback from the client as how accurate the prediction was by the application and the correction could be updated and the dataset could be improved and through consulting a real doctor after using the application could improve the application to the point that consulting a real doctor would become a rare case and most of the diseases could be predicted by the application.



IX. CONCLUSION AND FUTURE SCOPE

The framework would definitely lessen the human exertion, lessen the cost and time imperative in terms of HR and mastery, and increment the symptomatic exactness. The forecast of illnesses utilizing Data Mining applications and some unsafe undertaking as the information found are unessential and monstrous as well. In this situation, learning of the medicinal information is possible through information mining devices which proven to be useful and it is very fascinating. The scope of this paper could be commercial use of the application or further research purposes such as to detect the location of users and estimate which disease is more prevalent in a particular region and also to get results month wise that the frequency of a particular disease is diagnosed the most and spread awareness according to it in that region.

This paper gave a diagram of utilization of information mining procedures in regulatory, clinical, inquire about, furthermore, instructive parts of Clinical Predictions. This paper set up that while the current down to earth utilization of information mining in wellbeing related issues is constrained, there exists an extraordinary potential for information mining systems to enhance different parts of Clinical Predictions. Besides, the inescapable ascent of clinical information will build the potential for information mining systems that enhances the quality and reduces cost of social insurance.

This system has large scope as it has the following features which are:

• Automation of Disease Diagnosis.
• Paper free work helping the environment.
• To increase the efficiency, accuracy for the patients to help them in future.
• Managing the information related to diseases.

REFERENCES

1. Mário W. L. Moreira, Joel J. P. C. Rodrigues, Antonio M. B. Oliveira, Ronaldo F. Ramos, Kashif Saleem, "A preeclampsia diagnosis approach using Bayesian networks", Communications (ICC) 2016 IEEE International Conference on, pp. 1-5, 2016, ISSN 1938-1883.
2. Ameera M. Almasoud, Hend S. Al-Khalifa, Abdulmalik Al-Salman, "Recent developments in data mining applications and techniques", Digital Information Management (ICDIM) 2015 Tenth International Conference on, pp. 36-42, 2015
3. K Sowjanya, Ayush Singhal, Chaitali Choudhary, "MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices", Advance Computing Conference (IACC) 2015 IEEE International, pp. 397-402, 2015.
4. Bharathan Venkatesh, Danasingh Asir Antony Gnana Singh, Epiphany Jebamalar Leavline, Advance in Intelligent Systems and Computing, vol. 517, pp. 633, 2017, ISSN 2194-5357, ISBN 978-981-10-3173-1.
5. Felix Gräßer, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder. 2018. Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. In Proceedings of the 2018 International Conference on Digital Health (DH '18). ACM, New York, NY, USA, 121-125.
6. Effieient Algorithms to find Frequent Itemset Using Data MiningSagar Bhise1, Prof. Sweta Kale International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056.
7. Text Data Mining of Care Life Log by the Level of Care Required Using KeyGraph Muneo Kushima, Kenji Araki, Tomoyoshi Yamazaki, Sanae Araki, Taisuke Ogawa, Noboru Sonehara Proceedings of the International MultiConference of Engineers and Computer Scientists 2017 Vol I, IMECS 2017, March 15 - 17, 2017, Hong Kong.
8. Chen, W., Guo, H., Zhang, F., Pu, X., and Liu, X. (2012). Mining Schema Matching Between Heterogeneous Databases. In Consumer Electronics, Communications and Networks (CECNet), 2012 2nd International Conference on, pages 1128–1131. IEEE.
9. Han, J. and Gao, J. (2009). Research Challenges for Data Mining in Science and Engineering. Next Generation of Data Mining, pages 1–18.
10. H. Abe, S. Hirano and S. Tsumoto, "Evaluation Temporal Models Based on Association Mining From Medical Documents," Japan Association for Medical Informatics, vol. 27, no. 1, pp. 33-38, 2007.
11. Sun, Y., Han, J., Yan, X., and Yu, P. S. (2012). Mining Knowledge from Interconnected Data: A Heterogeneous Information Network Analysis Approach. Proceedings of the VLDB Endowment.
12. M. Akhil jabbar & Dr. Priti Chandrab "Heart Disease Prediction System using Associative Classification and Genetic Algorithm" International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT, 2012.
13. Nikhil N. Salvithal" Appraisal Management System using Data mining "International Journal of Computer Applications (0975 – 8887) Volume 135 – No.12, February 2016.
14. Tanvi Sharma, Anand Sharma & Vibhakar Mansotra "Performance Analysis of Data Mining Classification Techniques on Public Health Care Data" International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 4, Issue 6, June 2016.
15. V.Krishnaiah , Dr.G.Narsimha, Dr.N.Subhash Chandra "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques"(IJCSIT)International Journal of Computer Science and Information Technologies, Vol. 4 (1) , 2013, 39 - 45.
16. Jaimini Majali, Rishikesh & Niranjan, Vinamra Phatak "Data Mining Techniques For Diagnosis And Prognosis Of Cancer" International Journal of Advanced Research in Computer and CommunicationEngineering Vol. 4, Issue 3, March 2015.
17. Data Mining Techniques to Predict Diabetes Influenced Kidney Disease Swaroopa Shastri, Surekha, SaritaInternational Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2017 IJSRCSEIT | Volume 2 | Issue 4 | ISSN : 2456-3307.
18. DeFronzo Ralph. From the triumvirate to the ominous octet: a new paradigm for the treatment of type 2 diabetes mellitus. Diabetes. 2009;58:773-95.

AUTHORS PROFILE

Vivek Mudaliar pursuing his undergraduate degreein Bachelors of Technology in the department of Information Technology at SRM Institute of Science and Technology Kattankulathur Campus.In his fourth year he has various certifications in data science and machine learning field as well as he has completed several projects in field of machine learning and neural network also has done various projects in kaggle which a website for machine learning enthusiasts.In all he is profound in android application development as he has an HPE certification in this field and has skill to build websites from scratch.But the most interested field for him is the one he is still pursuing the field of data science.

Mr. P.Savaridaasan is a Assistant Professor in Information Technology Department at SRM Institute of Science and Technology at SRM Institute of Science and Technology.He has Nine Academic years of edifying experience in Teaching and develop learners to update in real-time knowledge and to excel them solution oriented.Three years of Industry experience in Technical aspects this comprises two years in deploying ERP products for business solutions and one year as Technician. He has interest in fields such as cloud computing, human computer interaction and digital forensics.One year of SAP Software Domain experience in ABAP/4. Have exposure in mentoring Research and Development projects in various domains.

Sachin Garg pursuing his undergraduate degree of Bachelors of Technology in the department of Information Technology at SRM Institute of Science and Technology Kattankulathur Campus. He has numerous certifications on machine learning and deep learning and has participated in multiple kaggle competitions. He has also finished below 1% leaderboard in many india-wide data science hackathons. His research interests include automatic machine learning, transfer learning, active learning , deep learning, hyperparameter optimization, computer vision, image analysis and pattern recognition.