International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

Earthquake Prediction using Machine Learning Algorithm

Pratiksha Bangar, Deeksha Gupta, Sonali Gaikwad, Bhagyashree Marekar, Jyoti Patil

Abstract: ***Per the statistics received from BBC, data varies for every earthquake occurred till date. Approximately, up to thousands are dead, about 50,000 are injured, around 1-3 Million are dislocated, while a significant amount go missing and homeless. Almost 100% structural damage is experienced. It also affects the economic loss, varying from 10 to 16 million dollars. A magnitude corresponding to 5 and above is classified as deadliest. The most life-threatening earthquake occurred till date took place in Indonesia where about 3 million were dead, 1-2 million were injured and the structural damage accounted to 100%. Hence, the consequences of earthquake are devastating and are not limited to loss and damage of living as well as non-living, but it also causes significant amount of change-from surrounding and lifestyle to economic. Every such parameter desiderates into forecasting earthquake. A couple of minutes' notice and individuals can act to shield themselves from damage and demise; can decrease harm and monetary misfortunes, and property, characteristic assets can be secured.***

In current scenario, an accurate forecaster is designed and developed, a system that will forecast the catastrophe. It focuses on detecting early signs of earthquake by using machine learning algorithms. System is entitled to basic steps of developing learning systems along with life cycle of data science. Data-sets for Indian sub-continental along with rest of the World are collected from government sources. Pre-processing of data is followed by construction of stacking model that combines Random Forest and Support Vector Machine Algorithms. Algorithms develop this mathematical model reliant on "training data-set". Model looks for pattern that leads to catastrophe and adapt to it in its building, so as to settle on choices and forecasts without being expressly customized to play out the task. After forecast, we broadcast the message to government officials and across various platforms.

The focus of information to obtain is keenly represented by the 3 factors – Time, Locality and Magnitude.

Keywords: Earthquake, Forecast, Machine Learning, Random Forest, Support vector Machine

Manuscript received on March 12, 2020.
Revised Manuscript received on March 25, 2020.
Manuscript published on March 30, 2020.

* Correspondence Author
Pratiksha Banagr*, Department of Information Technology, Jayawantrao Sawant College Of Engineering, Pune, India
Email: bangar.n.pratiksha14061998@gmail.com
Deeksha Gupta, Department Of Information Technology, Jayawantrao Sawant College Of Engineering, Pune, India
Email: deeksha.v.gupta27041998@gmail.com
Sonali Gaikwad, Department Of Information Technology, Jayawantrao Sawant College Of Engineering, Pune, India
Email: sonali.gaikwad04081998@gmail.com
Bhagyashree Marekar, Department Of Information Technology, Jayawantrao Sawant College Of Engineering, Pune, India
Email: bhagyashree.s.marekar01011999@gmail.com
Jyoti Patil, Ph. D. Research Scholar, Department of CSE, Koneru Lakshmaiah Education Foundation (KLEF), Guntur, A.P.India.
Email: jyotipatilnba@gmail.com

I. INTRODUCTION

Earthquake's association with structural damage and loss of life is one that keeps on enduring and thus is focal point of consideration for a many fields, say, seismological research and environmental engineering yet not limited to these[1]. It's significance is stretched out to human life too, for to sustain and to survive. A prediction that can be accurate and relied on is a requisite for all the areas prone to disasters and as well as for locations that have less to none chances. It will get us ready for all the worst possible scenarios and for necessary measures as well that can be taken before hand to solve upcoming crisis. As the technology is evolving and helping humans for a better and a convenient lifestyle, possibility at saving life is taken up with the help of efficient ML algorithm and Data Science to give accurate forecast. Machine Learning is a subset of Artificial Intelligence. It permits the system to adapt to a behaviour of a particular kind based on its own learning and possesses the ability to improve itself naturally solely from experience without any explicit programming, human mediation or help[8]. Initialisation of a machine learning process starts with feeding an honest quality data-set to the algorithm(s), so as to build a ML prediction model. Algorithms perform knowledge discovery and statistical evaluation, determining patterns and trends in data. Selection of algorithms relies on data and on the task that requires automation.

Our target is foreseeing catastrophic events and improving the manner in which we react to them. Great forecasts and admonitions spare lives. A notice of an approaching calamity can be issued well ahead of time as it will help in reducing both death occurrence and structural loss.

ML algorithms construct two types of predictive models, Regression and Classification models[6]. Each of them approaches data in a different way. Concerned system makes use of regression model whose core idea is forecasting a numerical value.

A. Earthquake Forecast

Anticipating a seismic event is viewed as an impossible phenomenon. It is a troublesome errand due to non-linearity of the event and unreliability [3] in it yet the ability of ML algorithms to assemble prescient models has transformed it into a potential wonder.

Earthquake forecast for Indian subcontinent along with rest of the World requires employing their earthquake catalogue aka data-set. A earthquake catalogue refers to a complete list of earthquake location, time, magnitude and depth that have happened in the past[3]. Methodology relies on sequence of these past earthquakes,
recognising suitable, necessary and appropriate parameters, identifying patterns in these parameters and understanding correlations between actual earthquakes from the past so as to predict future occurrence.

Various Random Forest-Support Vector Machine ensemble model are studied, modelled and deployed.

II. RELATED WORK

Wenrui Li , Nakshatra, Nishita Narvekar, Nitisha Raut, Birsen Sirkeci, Jerry Gao introduce us to the idea that a strong earthquake is followed by aftershocks. We can detect location of these aftershocks by analysis of arrival time of P-waves and S-waves. Data collection from 16 earthquake stations in SAC file format, which contains time series data and is a waveform, used by authors to study trends in P-waveand S-wave. Data is clipped followed by noise removal to only obtain needed waveform by means of triggering algorithm and filters. AR picker algorithm used to determine values of P-wave and S-wave arrival time which are treated as extracted feature. Waveform is then converted into ASCII format. Data is then fed to different machine learning models-SVM, Decision trees Random forest and linear regression for comparison purpose. Random Forest distinguishes between earthquake leading and non-earthquake leading data the best, with an accuracy of 90. Use of triangulation technique to calculate epicentre, predict arrival time of P-wave and S-wave and the difference between the two arrivals.[2]

Khawaja Muhammad Asim, Adnan Idris, Francisco Mart´ınez-A´ lvarez, Talat Iqbal carried out prediction of earthquake for Hindu-Kush region where small to medium earthquakes hit regularly, in accordance with tree based ensemble classifiers like rotboost, random forest and rotation forest. They employ earthquake data-set, and convert magnitude into binary classes, hence adapting concept of binary classification. A new combination of features based on 3 factors- Gutenberg-Richter relationship, seismic rate changes and distribution of fore-shock frequency. Highlighting factor is calculation of 51 seismic feature using suitable procedures and techniques. Since all the models performed exceptionally well, we can conclude the strategy of calculating 51 features was very effective. Rotation forest gives an accuracy of 95.9% and titles itself the best among rest models.[3] The useful insights for us come in the fact that for every region on this earth, a prediction model needs to be deployed however there is no prediction of when and of what magnitude will an earthquake occur of.

G.T Prasanna Kumari develops a classification model using ensemble learning methods. Emphasis is hugely on two notable ensemble algorithms, named Bagging and Boosting to foresee how creation of diverse ensembles improves precision of algorithm and how they contrast in their effectiveness with respect to traditional approach of constructing a single model, usually followed in ML to build classifiers. Bagging and Boosting are discussed in depth by specifying how each algorithm's process flow is different from the other, different ways in which they can be applied, their respective algorithms, powerfulness, achievements and limitations as well. She further discusses how performance of each differs for batch processing (data given at once) and online processing (data generation in a continuous manner). She concludes that ensembles are usually considered impractical for systems where online processing takes place but here, its performance is better than batch processing with an advantage of low run time, especially for larger data-sets.[4] Her insights are helpful for us in constructing our own ensemble models.

Ant´onioE Ruano, Maria G. Ruano, Pedro M. Ferreira, Ozias Barros, G.Madureira, Hamid R.Khosravani acquire seismic information from the PVAQ and the PESTR station of the seismic monitoring system. They mention a significant objective fact that detectors already present at such stations produce enormous number of bogus alarms and fail in detection of the event due to their being based upon a standard STA/LTA ratio. Thus they present a new seismic detector entitled to SVM classifier and its application is in a continuous manner on such stations. They compare specificity and recall measures obtained for each station, and conclude that the SVM classifier could differentiate between noise and seismic events successfully. Next, they shift their focus in reducing detection time in Early Warning System. Obtained results (88 and 110 sec) are too huge to be considered for deployment, so a new approach is inherited of overlapping windows and as a result, time obtained was 1.3 sec and 1.8 sec respectively. On the other hand, a change in values of recall and specificity, result in increase in correct detection and in false alarms as well.[5]

III. PROPOSED WORK

Developing predictive modelling involves gradual procedure. Tools which are conventionally used for developing model are Python, Hadoop and R.
Various steps involved are:

A. DATA ACQUISITION

Data acquisition is the process for bringing data for production use either from source outside the system and into the system, or from data produced by the system. This is the underlying advance to start and alludes to gathering required information. We obtain required data sets from government provided website such as –

- USGS.gov (United States Geological Survey)- Scientific agency of the United States government.[13]
- IMD.gov (India Meteorological Department)- Agency of the Ministry of Earth Sciences of the Government of India.[14]

Google Acquired Kaggle contains data-set collected from different agencies of different governments.

The columns in the data-set are -

- Date
- Time
- Latitude
- Longitude
- Magnitude
- Depth
-

primary data into a clean data set to make it suitable for use. It consists of two steps:

- Data Engineering
- Feature Engineering

Data Engineering

Real-World Data is not in a structured and compatible form, a per-cent of it could be found as incorrect, invalid, out-of-range, off-base, impossible as well as missing data which influence the outcomes causing them to be deceiving, misleading and incorrect. Irrelevant and unreliable data can make pattern recognition and knowledge discovery in the training phase progressively troublesome. Hence, it is the most significant advance in an ML framework and one needs to clean the information to dispose of such qualities or validate/correct them. It involves data integration, computing missing values, taking care of categorical values, transformation, and error correction.

Feature Engineering

It involves either Feature Selection or Feature Extraction and Feature Scaling.

A data set contains numerous of features which are random and may not be useful in prediction. Feature Engineering deals with reduction of random features under consideration and obtaining a set of minimum features which contribute to accurate prediction. Many algorithms are provided by ML for feature selection/extraction. Feature scaling is strategy used to standardize or normalize the range of features in the data-set. Feature Engineering is useful as it compresses the data, reduces the storage space, computation time and removes redundant features.

C. MODEL BUILDING

The yield of an ML algorithm is a 'model'. To begin with, the target variable and feature variable are comprehended and fetched. Second, the data-set is partitioned into training and testing data-set and third, the regressor/classifier model is constructed and fitted to training data-set.

In python, scikit-learn is a simple, basic, efficient open source library that executes a range of machine learning algorithms featuring various classification, regression and clustering algorithms using a unified interface.[15] Step by step building is as follows:

Building A Random Forest Regression Model :

Random forests are an ensemble learning method that can be fabricated for both regression as well as classification chore. It takes on the task of constructing multiple of decision trees during training and outputs the class that is mean prediction (regression) of each individual tree or the mode of the classes (classification). This huge number of trees represents a forest. Decision trees are rule based models; on a given training data-set with targets and features, the decision tree algorithm will come up with rules to carry out classification and regression. Features will be nodes and their presence and absence will represent likeliness. This helps in constructing a path of rules to work with. The root and splitting node is based on information gain or gini index[9]. In *Random Forest*, the root and splitting nodes are calculated in a random manner[9].

B. DATA PRE-PROCESSING

Data Pre-processing is a technique that converts given

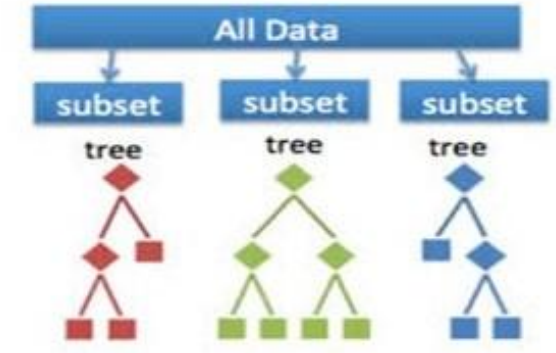

Fig. 1. Random Forest

Therefore random forest is a model comprising of various trees with the capability of making decision based on rule and the procedure of choosing root nodes and parent nodes is random.

Building A Support Vector Machine Regression Model :

Regression and classification chores can be performed by Support Vector Machines, a supervised learning algorithm. SVM segregates different data classes using a decision line named hyperplane. When predicting a numerical value, SVR attempts to find a function *f(x)* in the form of decision boundary at a certain deviation from €, which is a threshold value for all prediction to be within, from obtained targets value *Yi*, the original hyperplane, such that data points are within the boundary line. This decision boundary is the Margin of tolerance - a boundary that allows errors under given range.[10][11][12]

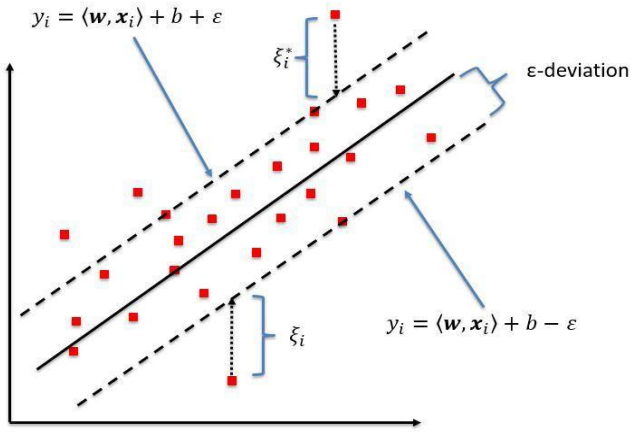

Fig. 2. Support Vector Regressor

Building A Stacking Regressor Model :

Stacking regression is an ensemble learning method. Several regression models collaborate, as a result, meta-regressor is build & itself finds its best fit by making use of output of individual regression models, trained on absolute training set, as meta-features.[7] Widely used to attain accuracy.

Fig 3, represents our model. "R1" and R2" are Random Forest and Support Vector Regressor respectively.

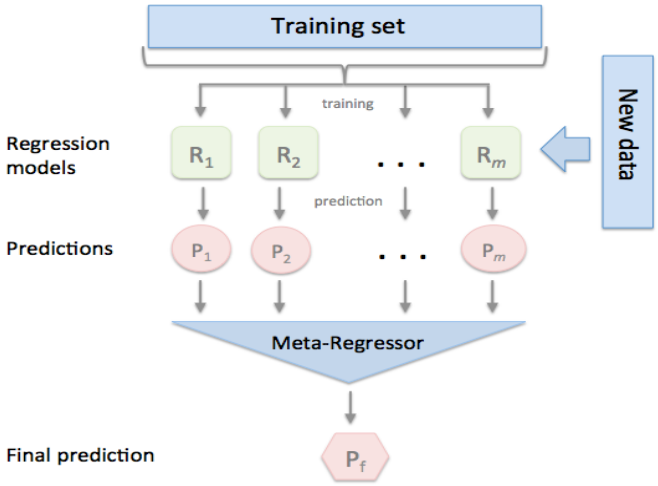

Fig. 3. Stacking

D. PREDICTIONS

Algorithm:

1.Input data-set and load libraries.
2.Data Pre-processing.
3.Model Building.
4.Making Predictions.

Data Visualisation:

1. Affected Areas

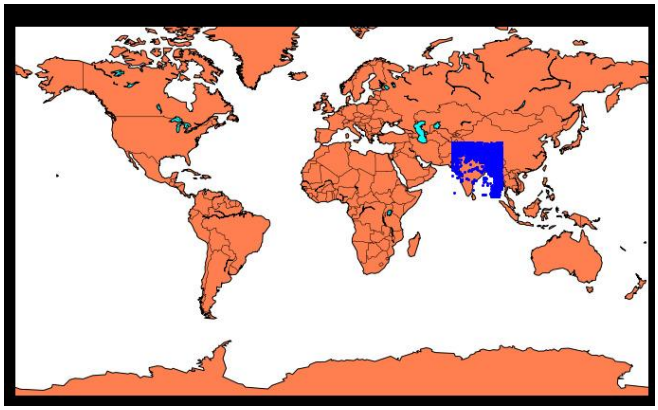

Fig. 4. Data Visualization for Indian Sub-Continent

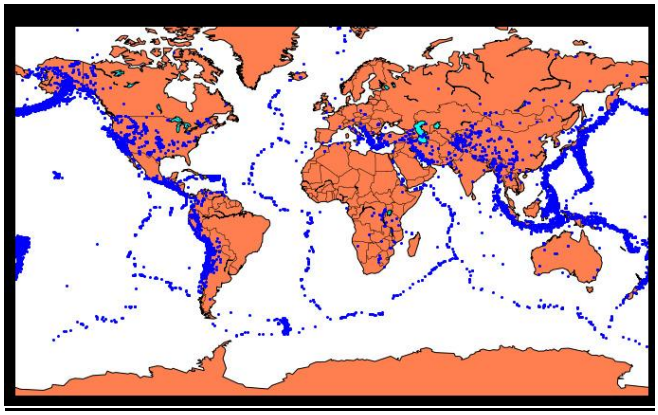

Fig. 5. Data Visualization for rest of The World

2. Prediction Using Bagging

Accuracy:74%

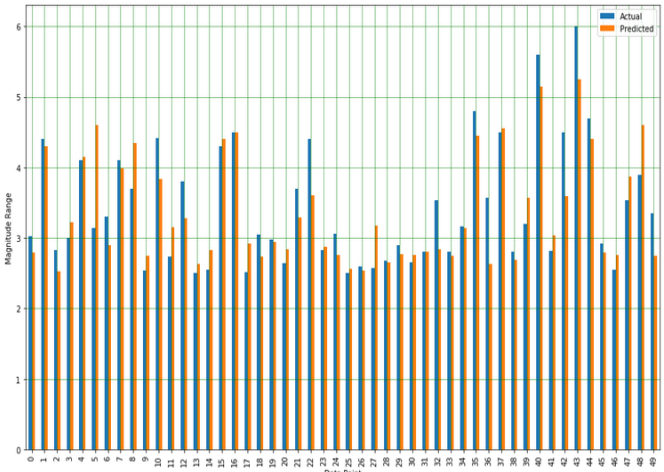

Fig. 6. Bar plot for Bagging

3. Prediction Using Boosting

Accuracy:76%

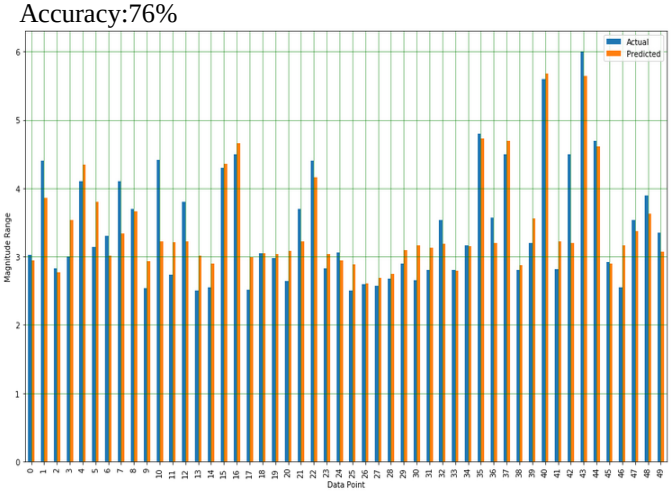

Fig. 6. Bar plot for Bagging

4. Prediction Using Stacking

Accuracy:83%

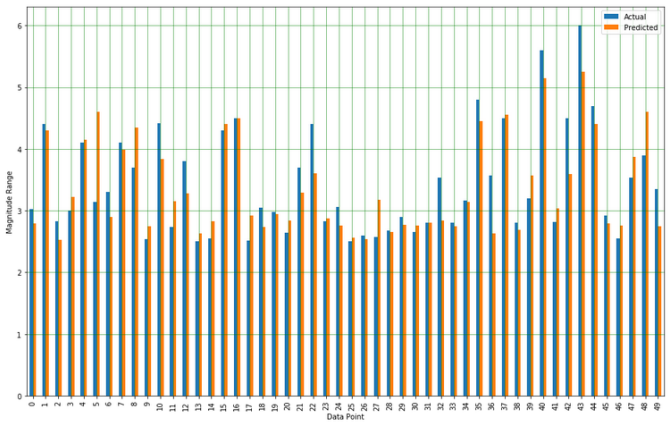

Fig. 6. Bar plot for Stacking

IV. RESULT

The randomforest-support vector machine model in combination work well for large dataset. The accuracy obtained for stacking model is the highest- 83% as compared to the accuracy of bagging and boosting. Response time is same for all the methodologies. Training time taken is slightly higher for stacking. Results are as follows :

Table- I: Result Table

Parametes/ Algorithms	ACCURACY	TRAINING TIME	RESPONSE TIME
Bagging	74%	3m5sec	5 sec
Boosting	76%	3m19sec	5sec
Stacking	83%	11m37sec	5sec

V. CONCLUSION

Thus we can conclude that integration of seismic activity with machine learning technology yields efficient and significant result and can be used to predict earthquakes widely, given the past history of the same is well maintained. Our attempt can be termed successful. The collaboration of the two can further be advanced to guard earthquakes more acutely. Large datasets prove to be very significant. Prediction models can be deployed in an area-centric manner, thus increasing the chances of accurate prediction exponentially but at the cost of studying algorithms used to build Stacking model, as it will perform well only if the algorithms chosen to build metaregressor are accurate themselves. The use of the methodology can be expanded in predicting various natural disasters as well.

REFERENCES

1. C. Li and X. Liu, ”An improved PSO-BP neural network and its application to earthquake prediction,” 2016 Chinese Control and Decision Conference (CCDC), Yinchuan, 2016, pp. 3434-3438.
2. W. Li, N. Narvekar, N. Nakshatra, N. Raut, B. Sirkeci and J. Gao, “Seismic Data Classification Using Machine Learning,” 2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService), Bamberg, 2018, pp. 56-63.
3. K. M. Asim, A. Idris, F. Mart´ınez-A´ lvarez and T. Iqbal, “Short Term Earthquake Prediction in Hindukush Region Using Tree Based Ensemble Learning,” 2016 International Conference on Frontiers of Information Technology (FIT), Islamabad, 2016, pp. 365-370.
4. Kumari, G. T. Prasanna. “A Study Of Bagging And Boosting Approaches To Develop Meta-Classifier.”, Engineering Science and Technology: An International Journal, Vol.2, 2012, pp. 850-855.
5. Ant´onio E Ruano, G. Madureira, Ozias Barros, Hamid R. Khosravani, Maria G. Ruano, Pedro M. Ferreira. ”A Support Vector Machine Seismic Detector for Early-Warning Applications”, IFAC Proceedings Volumes, 2013, pp. 400-405
6. Nick Minaie. “A Beginner’s Guide to Selecting Machine Learning Predictive Models in Python”, Towardsdatascience, Medium, 16 July 2019.
7. U. Pasupulety, A. Abdullah Anees, S. Anmol and B. R. Mohan, “Predicting Stock Prices using Ensemble Learning and Sentiment Analysis,” 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), Sardinia, Italy, 2019, pp. 215-222.
8. The Expert Team. “What is Machine Learning? A definition”, Expertsystem, 7 March 2017.
9. Flach, Peter. Machine Learning: the Art and Science of Algorithms That Make Sense of Data. Cambridge University Press, 2017, pp. 331-333
10. Support Vector Regression, www.saedsayad.com/supportvectormachinereg.htm.
11. Martin M. (2002) On-Line Support Vector Machine Regression. In:Elomaa T., Mannila H., Toivonen H. (eds) Machine Learning: ECML 2002. ECML 2002. Lecture Notes in Computer Science, vol 2430. Springer, Berlin, Heidelberg
12. Smola, A.J., Sch¨olkopf, B. A tutorial on support vector regression. Statistics and Computing 14, 199–222 (2004).
13. “United States Geological Survey.” Wikipedia, WikiMedia Foundation, 11 Mar. 2020, en.wikipedia.org/wiki/UnitedStatesGeologicalSurvey.
14. “India Meteorological Department.” Wikipedia, Wikimedia Foundation, 21 Jan. 2020, en.wikipedia.org/wiki/IndiaMeteorologicalDepartment.
15. Kumar, Vivek. “Vivek Kumar.” Pluralsight, 13 May 2019, www.pluralsight.com/guides/building-classification-models-scikit-learn

AUTHORS PROFILE

Pratiksha Bangar, an undergraduate, is pursuing Bachelor of Engineering, in the branch of Information Technology from Department Of Information Technology, JSPM’s Jaywantrao Sawant College of Engineering, Pune and currently is in her final year. Research area is Machine Learning.

Deeksha Gupta, an undergraduate, is pursuing Bachelor of Engineering, in the branch of Information Technology from Department Of Information Technology, JSPM’s Jaywantrao Sawant College of Engineering, Pune and currently is in her final year. Area of Interest is Machine Learning.

Sonali Gaikwad, an undergraduate, is pursuing Bachelor of Engineering, in the branch of Information Technology from Department Of Information Technology, JSPM’s Jaywantrao Sawant College of Engineering, Pune and currently is in her final year. Research area is Machine Learning.

Bhagyashree Marekar, an undergraduate is pursuing Bachelor of Engineering, in the branch of Information Technology from Department Of Information Technology, JSPM’s Jaywantrao Sawant College of Engineering, Pune and currently is in her final year. Area of interest is Machine Learning.

MS. Jyoti Patil, Currently working as Head of Department and Associate professor Department of Information Technology JSPM’s Jayawantrao Sawant college of Engineering Hadapsar ,Pune,India. Her Major Area of Research are Deep learning,Data Analytics, Hadoop MapReduce, Image processing. She is Pursuing Ph.D. in deep learning bigdata from KLEF deemed to be University, Guntur,A.P,India.She has published almost 11 papers in national and international journals.She has Published Patent On “Detection Of Brain Tumor Levels In MapReduce 3D MRI Images Using Hadoop”