International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2063-2069 www.ijaem.net ISSN: 2395-5252

"A Brief Study on Random Forest Using Python"

Madhav,Shubham Kohli, HimanshuRawat, Priyanshu Joshi
Department of Information Technology, Dr. Akhilesh Das Gupta Institute of Technology and Management, New Delhi

Submitted: 05-06-2021 Revised: 18-06-2021 Accepted: 20-06-2021

ABSTRACT— The goal of the project was to predict whether or not a SOS call for disaster happened on a particular day. We have downloaded data from the organization's internal database and removed some information for privacy. Loading the data into python and pandas allows for some simple plots to make sure we're focusing on the most relevant data.
Some exploratory plots show that we have calls on 29% of days, which means the data is somewhat imbalanced, but not horribly so. Calls happen most frequently on weekends and in the summer. We know this by intuition, and the data backs this up. This also confirms that there is probably some information within the date alone that may have some predictive power. There are several data sources we're using, and we'll be linking all of them together by the date. Python's date-time library was key for this and made these operations much easier

I. INTRODUCTION

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

Mathematical concepts behind Decision Tree

I. Entropy

Entropy is a measure of the randomness of a system. The entropy of sample space S is the expected number of bits needed to encode the class of a randomly drawn member of S. Here we have 14 rows in our data so 14 members.
Entropy $E(S)= -\sum p(x)*\log2(p(x))$

II. Information Gain

The information gain is the amount by which the Entropy of the system reduces due to the split that we have done.
$IG(S, a) = H(S) – H(S \mid a)$

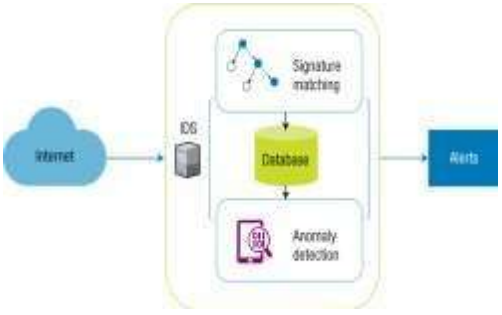

Figure : 1 DISASTER in IOT

1.1 Hyperparameter tuning:

A Machine Learning model is defined as a mathematical model with a number of parameters that need to be learned from the data. By training a model with existing data, we are able to fit the model parameters.
However, there is another kind of parameters, known as Hyperparameters, that cannot be directly

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2063-2069 www.ijaem.net ISSN: 2395-5252

learned from the regular training process. They are usually fixed before the actual training process begins. These parameters express important properties of the model such as its complexity or how fast it should learn.

Some examples of model hyperparameters include:

The penalty in Logistic Regression Classifier i.e. L1 or L2 regularization

The learning rate for training a neural network.

The C and sigma hyperparameters for support vector machines.

The k in k-nearest neighbors.

GridSearch CV

In grid Search CV approach, machine learning model is evaluated for a range of hyperparameter values. This approach is called grid Search CV, because it searches for best set of hyperparameters from a grid of hyperparameters values.

RandomizedSearchCV

RandomizedSearchCV solves the drawbacks of grid Search CV, as it goes through only a fixed number of hyperparameter settings. It moves within the grid in random fashion to find the best set hyperparameters. This approach reduces unnecessary computation

Hyperparameter tuning requires explicit communication between the AI Platform Training training service and your training application. Your training application defines all the information that your model needs. You must define the hyperparameters (variables) that you want to adjust, and a target value for each hyperparameter.

II. LITERATUREREVIEW

In view of the fact that IoT represents a new concept for the Internet and smart data, itis a challenging area in thefield of computer science. The importantchallenges for researchers with respect to IoT consist of preparing andprocessing data and discovering knowledge.

In this research paper [2] the authors have used machine learning techniques, approaches or methods for securing things in IOT environment. This paper attempts to review the related research on machine learning approaches to secure IOT devices.

In this research [4] the various machine learningmethods that deal with the challenges presented by IOT data by considering smart cities as the main use case. Thekey contribution of this study is the presentation of taxonomy of machine learning algorithms explaining howdifferent techniques are applied to the data in order to extract higher level information. The potential andchallenges of machine learning for IOT data analytics will also be discussed. A use case of applying a SupportVector Machine (SVM) to Aarhus smart city traffic data is presented for a more detailed exploration.

In this research paper [5] authorsaims to provide a brief overview of machine learning methods for internet of things (IOT). Authors present some of the applications of machine learning in IOT and have tried to provide an overview of the types of ML, ML task and its applications as related to IoT. In conclusion, it is needful to mention that ML provides higher precision in calculations and for prediction, it is highly effective and is able to look at a lot of information in smaller interims of time.

In the research paper [8] authors review ML/DLmethods for IoT security and present the opportunities,advantages and shortcomings of each method. Authors discuss theopportunities and challenges involved in applying ML/DL to IoTsecurity. These opportunities and challenges can serve as potentialfuture researchdirections.

This research paper [9] addresses the comparison of severalfrequently used ML classifiers from the group of SVMlike classifiers, namely SMO and C-SMV algorithm, and arange of ensemble algorithms on the other side, namelyLAD Tree, REPTree, RF and Multi -Boost. The analysis isbased on a range of testing procedures in Weka, with agoal to estimate a set of selected performance metrics andmake classifier comparison. As the analysed UNSWNB15dataset belongs to a unbalanced dataset category,for the proper examination of the classifiers we haveassumed the need for calculating the precision, recall,ROC and necessary time for classification.

III. MACHINE LEARNING ENSEMBLE APPROACH BASED ON RANDOM FOREST FOR RISK EVALUATION OF REGIONAL Flood Disaster

The concept of risks has been around for a long time. The risk depends on the probability of occurrence and the outcome [7]. However, the best solution to flood disasters was to control them, until the late 20th century [8,9,10]. The interaction between human society and the ecological environment has become more and more profound with the development of urbanization. The United Nations [11] raised the concept of sustainable development in the 1980s. Since then, the issue of the socio-economic ecosystem has attracted the attention of scholars at home and abroad

In the late 1980s, the establishment of the

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2063-2069 www.ijaem.net ISSN: 2395-5252

Intergovernmental Panel on Climate Change (IPCC) marked the beginningofmodernfloodriskmanagement

IV. MACHINELEARNING

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data. Machine learning techniques have ability to implement a system that can learn from data. For example, a machine learning system could be trained on incoming packets to learn to distinguish between intrusive and normal packet. After learning, it can then be used to classify new incoming packets into intrusive and normal packets. In machine learning, computer algorithms (learners) attempt to automatically distill knowledge from example data. This knowledge can be used to make predictions about novel data in the future and to provide insight into the nature of the target concepts applied to the research at hand, this means that a computer would learn to classify alerts into incidents and non-incidents task. A possible performance measure (P) for this task would be the Accuracy with which the machine learning program classifies the instances correctly. Machine learning often included in the category of predictive analytics as it helps to predict the future analysis.

4.1. Types of Machine Learning

ML mainlydivided into three categories. Supervised and unsupervised arewidely used categories. In supervised machine algorithm, training data has input and its corresponding output.Unsupervised machine learning, we do not have any output.In reinforcement machine learning a software agentautomatic take action to maximize the performance or award. For active learning type, a PC can simply get information for a confined game plan of cases. Exactly when used instinctively, this information can be shown to the customer.

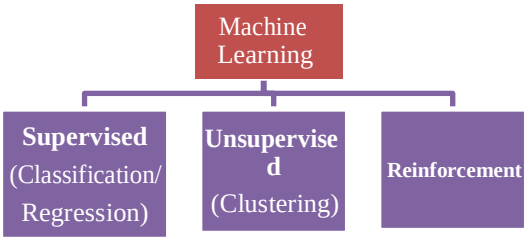

Figure 3: Types of Machine Learning

- **Supervised learning :**In this type of learning, theoutput class labels of the data are known or can be calculated. In cases where the labels are unknown,their operational data will be available.
- **Unsupervised learning:** No imprints, labelling orcategorization are given to the learning computation.It isolates the information to find the structure initsdata.
- **Reinforcement learning**: It is an area of Machine Learning. It is about taking suitable action to maximize reward in a particular situation.

4.2 Real Machine learningworkflow

- GatheringData
- Cleaningdata
- Model building and choosing the correct calculation
- Gaining insights from theoutcomes
- Visualizing theinformation

4.3 Machine LearningAlgorithms:-

A. Random Forest

Random Forest classifier infers that for a given class,features are independent [12]. Using the most frequent values of the features Random Forest classifier dispense theclass label to the instances [13]. It calculates the priorprobability of each class in the training phase using theoccurrences of the each feature for each class. Random forest finds the posterior probability of the class based onthe class prior probability [14]. It deduce that the result ofthe predictor for a given class is independent of the valuesof other predictor. Using the aforementioned probabilitiesit assigns the class label to the new data.

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2063-2069 www.ijaem.net ISSN: 2395-5252

B. Support Vector Machines(SVM)

SVM is a supervised ML algorithm with low computational complexity, used for classification and regression. It has the ability to work withbinary as well as with multi-class environments. It classifiesinput data into n dimensional space and draws n − 1 hyperplane todivide the entire data points intogroups.

C. J.48

J.48 is a type of decision tree. Decision tree considers theclass as a dependent variable which lies on the leaf of atree. Decision tree is a graphical representation of theclassification algorithm [15]. J.48 creates, first, a decisiontree in order to classify new instances. Dependentvariables (classes) are decided by the values of theinternal nodes which represent the variables which areconsidered independent variable.

4. Disater Dataset

The goal of the project was to predict whether or not a SAR call happened on a particular day. I've downloaded this from the organization's internal database and removed some information for privacy. Loading the data into python and pandas allows for some simple plots to make sure we're focusing on the most relevant data. I had to narrow the dates into a range from 2002 to current date, since the data before this was incomplete.
Some exploratory plots show that we have calls on 29% of days, which means the data is somewhat imbalanced, but not horribly so. Calls happen most frequently on weekends and in the summer.
We know this by intuition, and the data backs this up. This also confirms that there is probably some information within the date alone that may have some predictive power.

A. LogProbability

A log probability is simply the logarithm of a probability. The use of log probabilities means representing probabilities in logarithmic space, instead of the standard [0, 1] interval. In most machine learning tasks we actually formulate some probability p which should be maximized, here we would optimize the log probability log(p) instead of the probability for class θ. The use of log probabilities determines better numerical stability, when the probabilities are close to each other and very small.

ex = y
loge (y) = x

Where x is probability. To get back the values of probability take log of y on base e.

V. PROPOSED WORK

Some of the researchers in the field of machine learning has addressed the strategy for improve the performance of ML classifier which is used in modern intrusion detection system. To classify abnormal behaviorand minimizing misclassification propose a classification framework based on new Random Forestalgorithm are proposed. The Proposed Random Forest algorithm is used the concept of log probability. Detail about the log probability discuss inIntroduction.
Proposed Random ForestAlgorithm: Old Random Forest Algorithm
Begin To get Class of specific Instance state Probabilities of Array size = n
(n = total number of classes in dataset) Loop For j=0 to n-1
For each class get value of probability and save in probability [j]
End For
Get no. of attributes Loop, While
Declare variable temp and max=0; Loop For j=0 to n-1
Get probability estimates of each attribute and product over of these with each class probabilities.
Get max of these probability obtained in previous step and store in array of probabilities.
Now get / Take log of probabilities and update in array of probabilities.
Take max value from array of log of probabilities
End For
End while

This is proposed new Random Forest algorithm which is used for improving the DISASTER performance. We used Disaster dataset. The first step is pre-processing in this step clean the raw data and get ready to processed now in attribute extraction steps select appropriate attribute from dataset. In the next step, we applied new Random Forest classification algorithm on training and testing dataset in order to classify normal and abnormal data and measure performance. This same process also applied for general random forestclassifier algorithms and compare result. Architecture of the proposed work are shown in figure 2. For experiment purpose weka 3.8 tool is used.

International journal of advances in engineering and management (IJAEM)
Volume 3, issue 6 June 2021, pp: 2063-2069 www.ijaem.net ISSN: 2395-5252

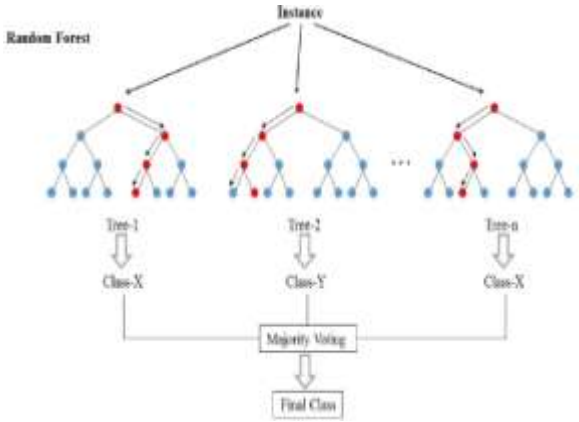

Figure 4: Proposed Classification framework

VI. RESULTANALYSIS:

The experimental result of the proposed Disaster model for flood risk shows that with proposed model random forest classifier gives better accuracy and take very less time to build model and improve the performance. Experimental result also compare with the performance of new Random Forest and general Random Forest algorithms. The performance parameter are as follow as: accuracy, error rate and time taken to build model. Table 1 show the comparison of experimental result. Result show that in comparison to general Random Forest multinomial text proposed new Random Forest algorithm give better accuracy and less error rate. time take to build model for new Random Forest is little bit max to general Random Forest.

Table 1: Comparison of result

Parameter	Random Forest Hyperparamter	General Random Forest
Accuracy	82.86 %	79.25 %
Error Rate	17.13 %	20.74 %
Time taken to build model	13.33 Second	0.14 Second

VII. CONCLUSION

Machine learning techniques are used for classification of data. Many existing study about the DISASTER are show that machine learning algorithms are used for classification of normal and abnormal data from large dataset. In this work new random forest classification algorithms based on hyperparameter is proposed. With the help of this random forest classifier, DISASTER improve the performance. Proposed work improves the performance of classifier which classifies the abnormal association, high accuracy and detection rate with low false alarm. The proposed work is completed by telling a framework for Classification and method to evaluate the framework. The issue of correct classification and model building time is also important for evaluating the framework. Proposed framework with new random forestclassification algorithms is showing greater accuracy when tested with general random forestclassifiers.

REFERENCES

[1]. D. Evans, "The Internet of Things How the Next Evolution of the Internet is Changing Everything," CISCO, 2011.

[2]. AmitSagu,NasibSinghGill–SecuringIoTEnvironment using Machine Learning Techniques‖ International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958,Volume-9 Issue-3, February, 2020

[3]. Chao Liang1, Bharanidharan Shanmugam1, Sami Azam1, Mirjam Jonkman1, Friso De Boer1, Ganthan Narayansamy2 "Intrusion Detection System for Internet of Things

based on a Machine Learning approach" 978-1-5386-9353-7/19/$31.00 ©2019 IEEE

[4]. YueXu–RecentMachineLearning Applications to Internet of Things (IoT)‖ Recent Machine Learning Applications to Internet of Things (IoT) Recent Machine Learning Applications to Internet of Things (IoT)

[5]. MohammadSaeidMahdavinejadMohammadrezaRezvanMohammadaminBarekatainPeymanAdibiPayamBarnaghiAmitP.Sheth[1,2][3] [4]–Machinelearningforinternetofthingsdataanalysis:asurvey‖http://www.keaipublishing.com/en/journals/digital-communications-and-networks/

[6]. Arun Kumar Rana1, AyodejiOlalekanSalau2,SwatiGupta3,SandeepArora4–A Survey of Machine Learning Methods for IoT and their Future Applications‖

[7]. Amity Journal of Computational Sciences (AJCS) Volume 2 Issue 2 ISSN: 2456-6616(Online)

[8]. Fei Wu, Limin Xiao, Jinbin Zhu "Bayesian Model Updating Method Based Android Malware Detection for IoT Services " 978-1-5386-7747-6/19/$31.00 ©2019 IEEE

[9]. Mohammed Ali Al-Garadi, Amr Mohamed, Abdulla Al-Ali, Xiaojiang Du, Mohsen Guizani–ASurveyofmachineanddeep learning methods for internet of things (IoT)Security‖

[10]. ValentinaTimcenko, SlavkoGajin "Machine learning based network anomaly detection for IoT environments" ieeeexplorer

[11]. Jadel Alsamiri1, Khalid Alsubhi2 "Internet of Things Cyber Attacks Detection using Machine Learning" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 12,2019

[12]. GiampaoloCasolla,SalvatoreCuomo,VincenzoSchiano di Cola , and Francesco Piccialli "Exploring Unsupervised Learning Techniques for the Internet of Things " 1551-3203 © 2019 IEEE

[13]. YU-XINMENG–ThePracticeonUsing Machine Learning For Network Anomaly Intrusion Detection‖ 2011 IEEE

[14]. Chi Cheng, Wee PengTay and Guang-Bin Huang–ExtremeLearningMachinesfor Intrusion Detection‖ - WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012 - Brisbane,Australia

[15]. NaeemSeliya , Taghi M. Khoshgoftaar–ActiveLearningwithNeuralNetworksfor Intrusion Detection‖ IEEE IRI 2010,August 4-6, 2010, Las Vegas, Nevada, USA 978-1- 4244-8099-9/10/$26.00 ©2010IEEE

[16]. KamarularifinAbdJalill, Mohamad NoormanMasrek–ComparisonofMachine Learning Algorithms Performance in Detecting Network Intrusion‖ 201O International Conference on Networking and Information Technology 978-1-4244-7578-0/$26.00 © 2010 IEEE

[17]. Shingo Mabu, Member, IEEE, Ci Chen,Nannan Lu, Kaoru Shimada, and Kotaro Hirasawa,Member,IEEE–AnIntrusion-Detection Model Based on FuzzyClass-Association-Rule Mining Using Genetic 739X/© 2019reserved. Elsevier B.V. All rights Network Programming‖ IEEE, JANUARY 2011

[18]. LiuHui,CAOYonghui–ResearchIntrusionDetection Techniquesfrom the Perspective of Machine Learning‖ - 2010 Second International Conference on MultiMedia and Information Technology 978-0-7695-4008-5/10 $26.00 © 2010 IEEE

[19]. Jingbo Yuan , Haixiao Li, Shunli Ding ,LiminCao–IntrusionDetectionModel

[20]. Maria Muntean, HonoriuVălean, LiviuMiclea,ArpadIncze–ANovel Intrusion Detection Method Based on Support Vector Machines‖ IEEE2010.

[21]. W. Yassin, Z. Muda, M.N. Sulaiman, N.I.Udzir,–IntrusionDetectionbasedonK-Means Clustering and OneRClassification‖ IEEE2011.

[22]. MohammadrezaEktefa, Sara Memar, Fatimah Sidi, Lilly SurianiAffendey–IntrusionDetectionUsing DataMining Techniques‖ IEEE 2010.

[23]. https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/bot_iot.php

[24]. HanwenWang,Biao Han, Jinshu Su," Biao Han, Jinshu S" 978-1-5386-9380-3/18/$31.00 ©2018 IEEE

[25]. IbraheemAljamal, Ali Tekeoglu Korkut Bekiroglu, Sangupta "Hybrid Intrusion Detection System Using Machine Learning Techniques in Cloud Computing Environment"978-1-7281-0798-1/19/$31.00©2019 IEEE SERA 2019, May

29-31,2019,Honolulu, Hawai

[26]. Fatima Hussain, Rasheed Hussain, Syed Ali Hassan, and EkramHossain "Machine Learning in IoT Security: Current Solutions and Future Challenges " arXiv: 1904. 05735v1 [cs.CR] 14 Mar 2019

[27]. NickolaosKoroniotis, NourMoustafa, Elena Sitnikova, BEnjamin Turnbull "Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset" https://doi.org/10.1016/j.future.2019.05.041 0167-