Proceedings of the International Conference on Smart Electronics and Communication (ICOSEC 2020)
IEEE Xplore Part Number: CFP20V90-ART; ISBN: 978-1-7281-5461-9

An Ensemble Approach to Predict Weather Forecast using Machine Learning

N.Sravanthi
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Guntur, India
E-mail - saisravanthi1808@gmail.com

M.Lohitha Venkat
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Guntur, India
E-mail – lohithavenkat28@gmail.com

S.Harshini
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Guntur, India
E-mail – harshinisaranu98@gmail.com

K.Ashesh
Assistant Professor
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Guntur, India
E-mail - imasheshk@gmail.com

***Abstract*— Weather changes have an incredibly negative impact on the environment and triggers natural disasters all of a sudden. To forecast these changes, there are several machine learning techniques and algorithms through which the weather changes can be predicted earlier. It has been noted that, from previous analysis there are many other approaches available for weather prediction. Based on those, various parameters like temperature, humidity, wind direction, precipitation, evaporation etc are considered. After surveying the emerging techniques and datasets, a proposed system is inculcated to include the approaches such as linear regression, bayes classifier, support vector machine and decision trees. In this the bagging, boosting, decision tree, random forest and stacking algorithms are used to predict the efficient accuracy. Bagging and boosting algorithms use same base learners whereas stacking uses different base learners. The learning capacity of stacking algorithm is different so that each individual learner can learn differently about various parameters and accuracy will increase when compared with other ensemble methods. Through the study it has been concluded to implement a proactive disaster recognition system to avoid the future loss of human lives and related environmental effect.**

Keywords: Prediction, Ensemble methods, Stacking, Bagging, Boosting, Framework.

I. INTRODUCTION

Artificial Intelligence [AI] is a man-made brainpower application, which gives frameworks the capacity to gain and improve naturally for a fact without being unequivocally customized. AI operates around the development of PC applications that can get knowledge and use it to develop about themselves. The study began, for example, with ideas or knowledge, models, direct understanding, or feedback, to look for designs in the knowledge and then decide on better choices based on the known models. The essential goal is to permit these PCs to consequently learn without human mediation or help, and to adjust activities in the similar manner.

Supervised Machine Learning Algorithm

It utilizes named guides to foresee future occasions, administered AI calculations will apply what has been realized in the past to new information. Starting with the investigation of a set up preparing dataset, the learning calculation creates a surmised capacity to make execution esteems forecasts. The framework can enable focuses to any new contribution after an adequate preparation.

Examples:

1. Characterizing spam emails
2. Labeling of web pages according to their content
3. Voice recognition

Unsupervised Machine Learning Algorithm These are utilized when preparing information is neither named nor grouped. Unaided learning investigates how frameworks can derive a capacity from unlabeled information to characterize a shrouded design. The framework doesn't discover the right yield, yet investigates the information and can attract surmising from datasets to clarify concealed structures from unlabeled information like clustering, visualization, dimensionality reduction.

Semi-Supervised Machine Learning Algorithm These calculations utilize both marked information and

unlabeled information to upgrade regulated learning. These arrangements with issues including a great deal of unlabeled information, and not very many marked information. It gives quicker and proficient execution in expectation in various fields and applications like chess game and object recognition.

A. PREDICTION

Prediction alludes to the presentation of a calculation once it has been prepared on a past dataset and executed on new information while anticipating the likelihood of a specific result, regarding whether a client would agitate inside 30 days [1]. For each record in the new information, the calculation will deliver likely qualities for an obscure characteristic, empowering the model manufacturer to figure out what that worth can most presumably be. Frequently it implies you're envisioning a potential outcome, similar to when you're utilizing AI to choose a promoting effort's next best move. Once in a while, for example, if an exchange that has just occurred was false or not.

R is a proposed in 1993 by Ross Ihaka and Robert Gentleman at the Auckland college and right now embraced by R advancement center group. It is a programming language utilized for factual figuring, information handling, graphical portrayal. It is a simple, all around created and proficient programming language that incorporates contingent articulations, circles, client characterized computational capacities and offices for info and yield. It is as often as possible used to create measurable programming

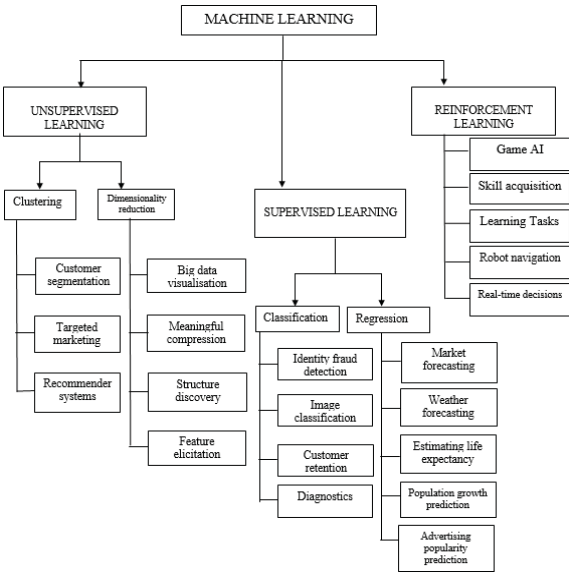

Figure 1: Machine learning classification

and information examination by analysts and information diggers. It has an effective stockpiling and information handling office. Some of the advantages of R programming are Open Source, Highly Compatible, Platform Independent, Quality Plotting and Graphing, Machine Learning Operations, Continuously Growing.

i. R Functions:

A huge amount of implicit component are available in

R coordinates the capacity contentions of your info parameters, either by esteem or by area, and afterward executes the body work. Capacity contentions may have default esteems: If such contentions are not indicated, R will take a default esteem. The source code for a capacity can be gotten to by running the capacity name itself in the debugger.

Customary capacities incorporate the capacities cbind (), rbind (), range (), sort (), order (). Every one of these capacities has a specific reason, and expects contentions to restore a yield. It is additionally required to create arbitrary information yet it required the numbers across machines to be comparative for learning and correlation. The set.seed () work is utilized with discretionary estimations of 123 to guarantee the most part that are producing similar outcomes. The quantity of components in a vector is returned by length R that has an assortment of numerical capacities and there is a wide assortment of measurable capacities in the ordinary establishment [10].

A client characterized work requires a name, contentions, and a structure. A smart thought is to call a client characterized work which is not quite the same as an incorporated capacity. Keeps away from vulnerability. The bits of code are reordered normally, more often than not Whenever RStudio is mentioned, R opens a situation. The available top-level condition is the worldwide framework, called R GlobalEnv.

ii. R in machine learning

The contrast between AI and information examination is somewhat mind boggling, yet the key idea is that AI seeks after prescient effectiveness over model interpretability, while information investigation favors interpretability and factual induction. R, as a language for factual surmising, effectively made its name in information examination. R has plentiful flexibility to do advance solid AI look into. R has bundles which upgrade its prescient precision. Caret is another bundle that bolsters R's AI capacities. It offers an assortment of capacities which improve prescient model advancement execution. As of recently, R has been utilized predominantly in the scholarly world and examination. Since R has been worked as a factual language, it has in general incredible measurable help. It mirrors the way wherein analysts think sensibly well, so it appears to be typical for somebody with a history in organized details. R makes exploratory work less complex than Python on the grounds that with just a couple of lines of code, measurable models can be composed. It is utilized to comprehend the working of furthermore applications dependent on calculations of Supervised and Unsupervised learning.

iii. Framework:

A structure is the development of a few prescient models for a provided dataset and after then coordinates these models into one last prescient framework. System is in any case called a gathering approach [1]. The main troupe technique appeared is sacking presented by Breiman dependent on bootstrap strategy. Later on, boosting was concocted dependent on Adaboost calculation by Freund and Schapire. Breimen additionally presented the irregular timberland calculation. These three calculations are in broadly use these days. These calculations are notable for their presentation since they more than once utilize single model to total the outcomes. Not just these standard models there are numerous different models where different grouping and expectation methods are utilized on similar information and can distinguish the best model among them. For instance, on the off chance to foresee climate on a specific day, parameters like stickiness, temperature and so forth will be viewed as it can utilize the choice tree for characterization and alongside this packing, boosting and arbitrary woods can be utilized at the same time among all these, one strategy gives ideal and exact outcomes.

iv. DATASET:

Based on the climatic conditions at our place, the parameters like temp, humidity, pressure, precipitation etc. are considered and a dataset is framed. These values are considered based on our city's weather situations which is Guntur, Andhra Pradesh and rest of the dataset values taken from the past data according to city's climatic condition. Around 1000 values are included in the dataset and designed dataset in such a way that there are few patterns in between the attributes and framed values according to the fluctuations of temperature and humidity.

II. METHODOLOGY

Methodology is the hypothetical, orderly research of the strategies applied to a zone of study. This incorporates the hypothetical investigation of the assortment of techniques and standards related with a surge of information. There are three kinds of strategies. They are: Explorative approach, Descriptive technique, Experimental philosophy. In our venture, the test approach is utilized.

A. Existing system:

Numerous examinations have created distinctive group strategies to enhance the expectation execution of the models. However, packing, boosting, and arbitrary woods are the most broadly utilized strategies [1]. Such procedures for the most part include visit utilization of a solitary model to join the outcomes. The model utilized is commonly the model of a choice tree.

i. Ensemble methodology:

It is a procedure of running various forecast models on a specific dataset and consolidates into a solitary last expectation model. The primary thought is that, an increasingly viable results are obtained when poor models are effectively consolidated. Most ordinarily utilized techniques are: Bagging, Boosting, Random woodlands, Stacking.

ii. Bagging:

Packing produces the best type of forecast by considering irregular information tests with substitution and of a similar size as that of the underlying dataset and regularly thinks about homogeneous students. At the point when a subset of the first information are taken, all the attributes are considered to parcel a hub. For each bootstrap information a model will be built for expectation, normal the models and experiences casting a ballot the most rehashed outcome will be the best one among all by decreasing forecast change and builds security.

iii. Boosting:

In producing numerous single models and totaling their presentation, boosting is like sacking and arbitrary woodland, yet changes from them in refreshing the loads of results at every emphasis while over and over utilizing same unique dataset. The point is to iteratively fit models so that model preparing at a given stage relies upon the models fitted in the past stages. "Boosting" is the most well-known of these techniques, as it creates an outfit strategy which is generally less one-sided than the powerless students that make it up.

iv. Random forest:

It is practically like packing, however the principle contrast is that in random forest, just a subset of highlights are picked indiscriminately from the aggregate and the best parting highlight from the subset is utilized to break every hub into a tree ,though in sacking all highlights are required to break a hub. They likewise vary in the expansion of irregular examining of factors during the bootstrap information age process. Random forest can diminish the expectation fluctuation in sacking by this arbitrary testing. Random forest can decrease the expectation change in packing by this arbitrary examining. The model f^((b)) (x) can be developed by random forest for every one of those bootstrap datasets and afterward at last produces a last model.

B. Proposed system:

Gathering approaches utilize diverse learning calculations to accomplish preferred prescient effectiveness over both of the separate learning calculations alone could likely accomplish. Aside from a statistical ensemble, which is typically unending in measurable mechanics, an AI outfit includes just a genuine limited scope of alternative models. Stacking includes showing a learning calculation (now and then called stacked speculation) to blend the expectations of numerous other learning calculations. Initial, a couple of different calculations are prepared with the current information and afterward a combiner calculation is prepared to make a last forecast utilizing all the past calculations' expectations as outer data sources. The fundamental detriment of expanding the levels is the scourge of information for the more elevated levels. As the levels are incremented, the informational collection continues partitioning and it can bring down the information for the more significant levels keeping the models from getting enough information to learn. In any case, the issue can be decreased by keeping up the legitimate exchange levels versus no. of models per level. Thinking about the stacking procedure as having two levels can be useful:

- Level 0: Level 0 information is the contributions of the

preparation dataset, and Level 0 models focus on making forecasts from this information.

• Level 1: Level 1 information utilizes the aftereffects of level 0 models as info and the level 1 model figures out how to reproduce this information.

Stacking strategy gives the accompanying advances: - 1.Divide the preparation information in 2 disjoint sets, Train various Base Learners in the initial segment, Check the Base Learners in the subsequent part and make expectations 4. Utilize the expectations from (3) as data sources, right yield reactions, train the Higher- Level Learner.

III. RESULTS

Ensemble approaches use different learning algorithms to achieve better predictive efficiency than either of the respective learning algorithms alone could probably achieve. The ensemble and predictive algorithms like Random forest, Linear discriminate analysis, generalized boosted algorithm, Support vector machine are utilized.

Linear Discriminant Analysis or LDA is a dimensionality decrease procedure. It is utilized as one of the pre-processing step in Machine Learning and utilizations of pattern alignment. The objective of LDA is to extend the highlights in high to a low-dimensional space so as to stay away from the scourge of dimensionality and furthermore lessen assets and dimensional expenses. . LDA centers basically around anticipating the highlights in higher measurement space to bring down measurements. it can be accomplished in three stages: Firstly, you have to ascertain the distinctness between classes which is the separation between the mean of various classes. This is known as the between-class change. Furthermore, ascertain the separation between the mean and test of each class. It is likewise called the within-class variance. At long last, build the lower- dimensional space which boosts between class change moreover limits the inside class fluctuation. P is considered as lower-dimensional space projection, which is otherwise called as Fisher's model.

Generalized Boosting Models fit various decision trees to enhance the exactness of the template. For each new tree in the model, an arbitrary subset of the considerable number of information is chosen utilizing the boosting . For each new tree in the model the information is weighted so the data which is worthless demonstrated by previous trees consists of higher probability of being chosen in the latest prototype. This implication follows the principal tree is suited, the model will contemplate the misunderstanding with the assumption of that tree to apt the following tree, etc. By taking into consideration of the attack of past trees that are fabricated, the framework ceaselessly strive to improve its veracity.

SVM is a managed AI computation which can be employed for either classification or regression challenges. In the SVM calculation, every datum is plotted as a point in n-dimensional space with the estimation of each element being the expectation of a particular organize. By then, the grouping is performed by detecting the hyper-plane that separates the two classes very well. Hyper-planes are decision confines that help portray the data centers. Data focuses falling around either side of the plane can be credited to different classes. Similarly, the part of the hyper-surface depends on the amount of features. In case the amount of data features is 2, by then the hyper-plane is just a line. If the amount of data features is 3, by then the hyper-surface moves toward a two-dimensional plane. It gets hard to imagine when the amount of features outperforms 3.

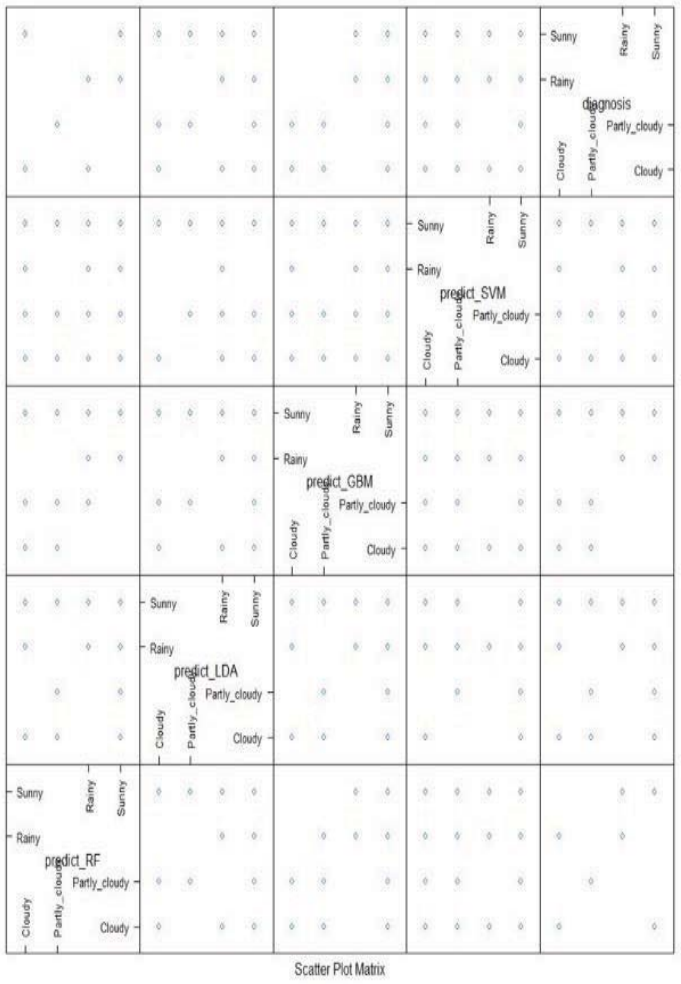

Figure 2: Scatter plot matrix

The above scatter plot depicts the performance of the various machine learning techniques. Among all the algorithm which is a framework after combining all the methods gives the more accurate results and Random forest gives less performance when compared to LDA, GBM, SVM.

Figure 3: Frequency of methods used

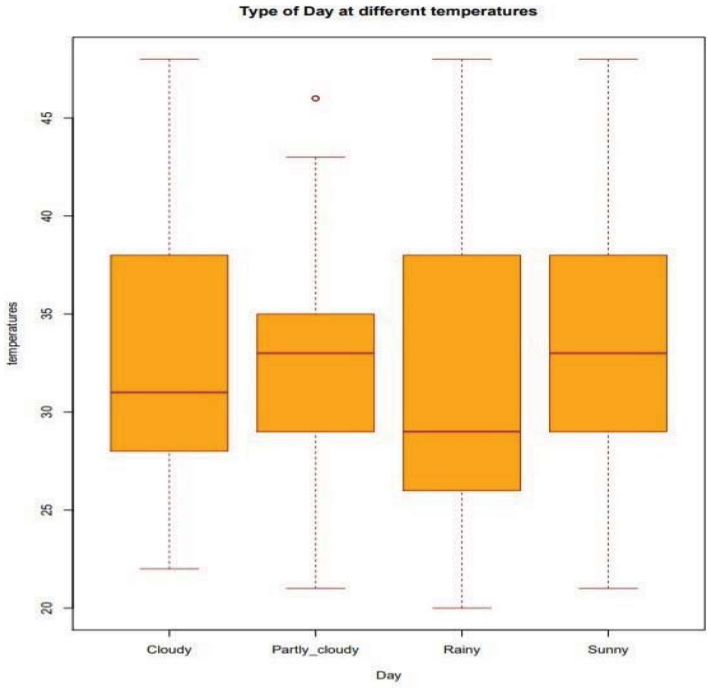

Figure 4: Box plot between temp versus day

The above figure represents temperature during day time when the weather may be cloudy, partly cloudy, rainy or sunny.

IV. CONCLUSION

The proposed research work has developed a model for weather prediction that can be utilized to provide better performance without much additional cost and also prediction variance can be reduced. Weather plays a major role in our daily life, and without the meteorologist and forecaster, it would have faced difficulty in planning the daily activities. Meteorologist and forecasters predict the weather and its possible changes, but in reality, weather is still unpredictable.

V. REFERENCES

[1] M. Ahmed, "A data fusion framework for real-time risk assessment on freeways," 2012.

[2] AHMJakaria, "Smart Weather Forecasting Using Machine Learning: A Case Study in Tennessee," 2018.

[3] O. Y. Al-Jarrah, "Efficient Machine Learning for Big Data: A Review," 2015.

[4] M. S. S. S. A. S. A. J. Amit Kumar Agarwal, "Forecasting using Machine Learning," 2019.

[5] B.Vasantha, "Rainfall Pattern Prediction Using Real Time Global Climate Parameters Through Machine Learning," 2019.

[6] C. Choi, "Development of Heavy Rain Damage Prediction Model Using Machine Learning Based on Big Data [Changhyun Choi] [2018]," 2018.

[7] Y. Di, "Prediction of Long-Lead Heavy Precipitation Events Aided by Machine Learning," 2015.

[8] C. Feng, "Hourly-Similarity Based Solar Forecasting Using Multi-Model Machine Learning Blending," 2018.

[9] GylianVerstraete, "A data-driven framework for predicting weather impact on high-volume low-margin retail products," 2018.

[10] N. D. Hoai, "Downscaling Global Weather Forecast Outputs Using ANN for Flood Prediction," 2011.

[11] A. Koesdwiady, "Improving Traffic Flow Prediction With Weather Information in Connected Cars: A Deep Learning Approach," 2016.

[12] J. Lee, "Constructing Efficient Regional Hazardous Weather Prediction Models through Big Data Analysis," 2016.

[13] S. Madan, "Analysis of Weather Prediction using Machine Learning & Big Data," p. 2018.

[14] M. C. R. Murça, "Identification, Characterization, and Prediction of Traffic Flow Patterns in Multi-Airport Systems," 2018.

[15] J. Wu, "Prediction of hourly solar radiation with multi-model framework," 2013.