

**School of Information Technology & Engineering**

**SWE 2009 – Data Mining Techniques**

**SLOT: A1+TA1**

**J Component**

**Review 2**

**Topic: Heavy Rainfall Prediction using Gini Index in**

**Decision Tree**

**Submitted by**

|  |  |  |
| --- | --- | --- |
| **S No** | **Reg No** | **Name** |
| **1** | **16MIS0121** | **Preethi Reddy** |
| **2** | **16MIS0457** | **Ayisha Siddiqua L** |
| **3** | **16MIS0468** | **Sri Likitha M L** |

**Faculty Incharge: Prof. Senthil Kumar N C**

**I. Detailed Literature Review**

[1] In this paper, a data-mining approach is applied to predict rainfall in a watershed basin at Oxford, Iowa, based on radar reflectivity and tipping-bucket (TB) data. Five data-mining algorithms, neural network, random forest, classification and regression tree, support vector machine, and *k*-nearest neighbor, are employed to build prediction models. The algorithm offering the highest accuracy is selected for further study. The construction of a predictive model is preceded by variable (input) selection aimed at the reduction of the dimensionality and complexity of the data set, as well as gains in model accuracy and computational efficiency. The proposed methodology has demonstrated high-accuracy rainfall predictions in Oxford, Iowa. It has provided useful hydrological information for SWAT models predicting water quality. In future research, data from other regions will be collected for further validation and improvement of the proposed approach.

[2] This research paper is motivated by the need to compare ANN (Artificial Neural Network) and FL (Fuzzy Logic) models to know which one is more efficient in predicting rainfall. The rainfall datasets used in this research work were collected from an automatic weather station in Iju, a town in Akure North Local Government Area of Ondo State for the period of four years (2007-2010). The model comparison is based on four criteria; the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), prediction error, and the prediction accuracy. The error measures are comparable for the two models. The analysis of the models accuracy, shows that, overall, the ANN model perform slightly better than the FL model in terms of PE, RMSE, MAE and accuracy. It is therefore recommended that the ANN and FL techniques could be improved upon by combining it with another method i.e., genetic algorithm for its optimization purpose.

[3] This paper presents a new approach using an Artificial Neural Network technique to improve rainfall forecast performance. A real world case study was set up in Bangkok; 4 years of hourly data from 75 rain gauge stations in the area were used to develop the ANN model. The developed ANN model is being applied for real time rainfall forecasting and flood management in Bangkok, Thailand. The common method used in the paper is trial and error based on a total error criterion. This method starts with a small number of nodes, gradually increasing the network size until the desired accuracy is achieved. Although the model performance of 6 hour forecasting was low and the forecasting was not as accurate as expected, the developed model can still be used for practical applications such as rainfall forecasting and flood management for the urban areas.

[4] This paper studies data intensive model using data mining technique to address the problem of compute intensive rainfall prediction model. This model works with efficient accuracy and uses moderate amount of compute resources for rainfall prediction. Bayesian approach is used for rainfall prediction. It works well with good accuracy. Weather data is one of the meteorological data that is rich with important information, which can be used for weather prediction. They extracted weather data which was collected from Indian Meteorological Department (IMD) Pune. From the collected weather data comprising of 36 attributes, only 7 attributes were most relevant to rainfall prediction. The model has simplicity, good prediction performance, and can be used for both binary and multiclass prediction problems. This model is nearly accurate model in comparison with compute intensive models. The performance of the model can also be improved by designing the model for scalable platforms, either for vertical scalability or for horizontal scalability.

[5] In this research paper, analysis of various popular data mining algorithms is presented for

rainfall prediction. Recent algorithms analyzed in this work are Naive Bayes, K- Nearest Neighbour algorithm, Decision Tree, Neural Network and Fuzzy logic are some of the algorithms compared in this paper.The data of monthly rainfall (in MM)for Coastal Andhra, Telangana and Rayalaseema regions in Andhra Pradesh state during the years 1871-2011 is collected from Climatology & Hydrometeorology Division, Indian Institute of Tropical Meteorology (IITM), Pune, India. The parameters include the Root Mean Square Error RMSE, Mean Absolute Error MAE, Coefficient of Correlation CC and BIAS. Two-Third of the data was used for training the model and One-third for testing. In the future works, some additional inputs were employed for rainfall prediction such as Sea Surface Temperature (SST) areas around Andhra Pradesh and Southern part of India.

[6] This research paper is focused on using the gini index as an attribute selection measure in an elegant decision tree to predict precipitation for voluminous datasets. This study aims at improving the prediction of precipitation over the supervised learning in a Quest decision tree, especially when the datasets are large. Four data sets have been used, out of which three datasets are of actual cities data. They have used monsoon period data. A decision tree using the gini index increases the accuracy rate while decreasing computational time by reducing the computation of total split points. This approach provides an average accuracy of 72.98% with a reduction of 63% in computational time over a SLIQ decision tree.

[7] The study in this paper assessed the ability of the Weather Research and Forecasting model to simulate rainfall over Western Uganda for the period 21st April to 10th May 2013 and tested six cumulus parameterization schemes. The root mean square error, mean error and the sign test method are used to assess the ability of the schemes to simulate rainfall along with an adapted contingency table. Results show that the Grell-Fretas scheme is better at simulating rainfall compared to other schemes over the study period while the Betts-Miller-Janji’c and the Kain-Fritsch schemes overestimated rainfall. However all the schemes under predicted heavy rainfall events but the Betts-Miller-Janjic and the Kain-Fritsch schemes over predicted the light rainfall. The variation of altitude presented a noticeable change in predicted rainfall where an increase of 25% in altitude increased the probability of prediction by 6.5% which shows a key role played by altitude in convection.

[8] This paper highlights a model using decision tree to predict weather phenomena like fog, rainfall, cyclones and thunderstorms, which can be a life saving information and used by peoples of all walks of life in making wise and intelligent decisions. The proposed model is implemented using the open source data mining tool Rapidminer. They mentioned notable decision-tree algorithms like ID3 (Iterative dichotomiser3), C4.5 (successor of ID3), CART, CHAID (CHI-squared Automatic Interaction Detector) and MARS (extends decision trees to better handle numerical data). This model may be used in machine learning and further promises the scope for improvement as more and more relevant attributes can be used in predicting the dependent variables.

[9] This paper presents a review of different rainfall prediction techniques for the early prediction of rainfall prediction of rainfall. Widely used techniques for prediction are Regression analysis, clustering, and Artificial Neural Network (ANN) etc. The Empirical approach is based on analysis of past historical data of weather and its relationship to a variety of atmospheric variables over different parts of Chhattisgarh. Some limitations is clearly noticed in all the methods of rainfall prediction discussed in this survey paper The extensive references in support of the different developments of methods provided in this research would be of great help to researchers to accurately predict rainfall in the future and to select the method that would solve their problem which they will be facing in their proposed prediction model.

[10] In this paper, they have developed a method to predict heavy rainfall in South Korea with a lead time of one to six hours. They have modified the AWS data for the recent four years to perform efficient prediction, through normalizing them to numeric values between 0 and 1and under sampling them by adjusting the sampling sizes of no-heavy-rain to be equal to the size of heavy-rain. Differential Evolutionary algorithms were used to select important features. Discriminant functions, such as support vector machine (SVM), k-nearest neighbors algorithm (k-NN), Normalization. In future work, they will pre-process the weather data by various methods, such as representation learning, cyclic loess, contrast, and quantile normalization algorithms.

[11] ANN has been successfully used by most of the researchers in this field for the last twenty-five years. Back propagation, is a common method of teaching artificial neural networks how to perform a given task. It is a supervised learning method, and is an implementation of the Delta rule. The significance of Back Propagation algorithm is that it can minimise the error in network by adjusting the weight through which the neurons are connected to each other. This paper reports a survey on rainfall predictions using BPN based neural network architectures. From the survey it has been found that most of the researchers used back propagation network for rainfall prediction and got significant results. The main advantage of the BPN neural network method is that it can fairly approximate a large class of functions.

[12] In this paper, analysis of various data mining algorithms is presented for rainfall prediction. Data Mining deploys techniques based on machine learning, alongside the conventional methods. More importantly, these techniques can generate decision or prediction models, based on historical data. Based on this analysis BP is combined with various other algorithms. Recent algorithms analyzed in this paper are ANFIS, ARIMA, SLIQ Decision Tree which used for prediction of Rainfall.

[13] Rainfall prediction is a beneficial but challenging task. Data mining techniques have the ability to predict the rainfall by extracting and using the hidden knowledge from past weather data. In the last decade, many researchers have worked to increase the accuracy of rainfall prediction by optimizing and integrating data mining techniques. Various models and techniques are available today for effective rainfall prediction but still there was a lack of a compact literature review and systematic mapping study which could reflect the current problems, proposed solutions and the latest trends in this domain. This research provided a comprehensive systematic mapping as well as the critical review of latest research from 2013 till 2017 in the area of rainfall prediction by focusing on data mining techniques. In this research a list of significant research questions was identified and then a systematic research process was followed to extract and shortlist the most relevant research articles from renowned digital search libraries. Answers of the identified questions were explored by critically reviewing the shortlisted articles. The research focus on the domain of rainfall prediction has been increasing since last decade and so are the problem areas. So it was concluded that enhancements, optimizations and integrations of data mining methods are vital to explore and solve these problems.

[14] Rainfall time series may be unfounded. The topic of monsoon-rainfall data series is highly complex; the role that multiple linear regressions might play in this topic is one for future research—it appears, from the evidence here, not to be useful as a predictive model. Whether it might be useful for offering an approximate value of future monsoon rainfall remains to be seen. Using this regression method, we have to forecast rainfall for our state also.

[15] From our observation, we come to a conclusion that there is a reduction in the rainfall for the next consecutive years starting from 2018 to 2020. It is also to be noted that amount of rainfall received is decreasing every year. Therefore this research has its significance in making the government organizations, NGOs and agriculturists living in this region to take precautionary measures to face the consequences that may arise because of this rainfall reduction. This research can be extended with more oceanic parameters, and other parameters like air temperature, surface temperature, and soil temperature etc. The other aspect of taking this research further is by comparing the results by adopting different methodology and algorithms like Clustering, ANN and Fuzzy logic. This will give a new vision for selecting the appropriate algorithm based on the different parameters that are considered for our prediction analysis.

[16] This research performed rainfall prediction in Lahore city using five data mining techniques: Support Vector Machine, Naïve Bayes, k Nearest Neighbor, Decision Tree and Multilayer Perceptron. 12 years of past weather data from December 1, 2005 to November 31, 2017, is used for prediction in this research. Performance analysis of used data mining techniques is performed using three accuracy measures: precision, recall and f-measure and results are presented in tables and graphs. For effective prediction, a classification framework is used in which the input data went through a pre-processing stage and got cleaned and normalized before classification process. To analyze the performance dependency of classification techniques on training data, ten ratios of training and test data (training data: test data) are used from 10:90 to 90:10. According to results, used classification techniques performed well for no-rain class however for rain class, the techniques did not perform well. The reasons behind the lower accuracy in rain class may include: missing values, absence of important climatic attributes in dataset and overall lower rate of rainfall in the city. It is suggested for future work that further predictions should be performed by exploring more classification techniques and climatic attributes on different weather data.

[17] Data mining approach for rainfall prediction model is data intensive model instead of compute intensive model which are being used in prediction centres. This model is nearly accurate model in comparison with compute intensive models. Because of data mining approach, computing power is reduced. The model returns good prediction results when the training dataset is large,. The negative part of model is, when a predictor category is not present in the training data, the model assumes that a new record with that category has zero probability. . The performance of the model can also be improved by designing the model for scalable platforms, either for vertical scalability or for horizontal scalability.

[18] During the simulation of the algorithms, in the graph some points were not forming cluster rather these points were away from the clusters and the straight line distance (Euclidean Distance) was greater. We faced this in some month’s data. As a result, the result was giving quite different value than the actual rainfall amount in that month. Otherwise the simulation gave us a closer value to the actual data. We hope this thesis will help people, especially the farmers who depend on the weather to grow their crops. They will come to know about the possibility of rainfall long before it will take place and a huge amount of disasters due to the heavy rainfall and flood will be shorten in future.

[19] In India, rainfall is a critical factor in farm management and water resource management. In this survey paper we found the use of various data mining techniques on the collected data set from the various resources may found useful in accurate prediction of rainfall. In this way Data mining offer us a much-needed opportunity to deliver scientific findings and information to stakeholders and decision makers for providing collective decision-making tools.It is observed from various studies that rainfall estimation and prediction varies from using MLR to SLIQ.

[20] Rainfall is the major cause for many of the natural disasters like flash floods, droughts, tsunamis. So in order to prevent these natural calamities, we should be able to predict the cause of the source. The proposed system can be used to estimate the rainfall over the required period so that the respective authorities can take precautions to prevent the loss of life and property. The proposed system uses modified linear regression approach to predict the rainfall that has less error percentage than compared to most data mining techniques like clustering, back propagation which provides the generalized values rather than estimate values. This data is used to perform the necessary calculations to predict the rainfall from average temperature and cloud cover of that particular district. We may improve this system further using multiple regression which can take multiple months at a time as input and just forming a single equation which leads nearer to an accurate rainfall predicted. The proposed approach may also be used in other applications like, in schools to predict the average marks of their students, in sports to predict the scores or winning teams based on their previous performance, in enterprises to estimate their profits, etc.,.

[21] This paper uses K means clustering technique, classification and regression Tree, Statistical downscaling method is used to predict the future rainfall places using GCM simulated climate variables. This model relates the large scale variables which are called as predictors to the local scale climate which is called as predictands. Transfer function based method is used in order to get the linear and non-linear regression methods between the predictors and predictands. Data is collected from National Centers for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) reanalysis climatic data set Since we are using clustering techniques it require many clusters as input which is major problem to overcome we use validation measures and they are Dunn’s index, e Davies–Bouldin index , and the Silhouette index which ensures compactness , connectedness and separation of cluster partitions.

[22] This paper uses Bayesian enhanced modified approach (BEMA), uses Artificial Neural Network, non linear filters are used. In order to improve the performance of BEMA filter that comes under Artifical neural network for comparision we use Bayesian Approach(BA). In this the computational result are is evaluated on high roughness time series and compared with artificial neural netwroks and with its non linear filters. In this particular paper prediction for complete and incomplete dataset is established and achieved using this BEMA combined with permutation entropy. The main key area of this paper is incomplete data or missing data done by changing the structure of the predictors based on the data model selected,In which the approach is combined along with the entropy information. The missing data in the dataset is imputed using linear average smoothening

[23] This paper uses Bayesian approach, Data intensive model. The performance of the model can also be improved by designing the model for scalable platforms, either for vertical scalability or for horizontal scalability. The training dataset is compared with the available dataset in order to get the accuracy. First the data undergoes pre -processing and later data transformation is done in order to work on Bayesian. Data set is collected which contains 36 attrbutes ut of which only 7 attributes are related rainfall prediction. First the data undergoes pre -processing and later data transformation is done in order to work on Bayesian

[24] This paper uses Wavelet technique, Artificial Neural Network. Wavelet analysis plays a major role when compared to all other methods like Fourier series. Wavelet concept can be applied to any size of time series. This is used to explore, de noise, smoothen time series which can be used for other empirical purposes. This study used only dataset from one particular gague

[25] This paper uses Regression, Artificial Neural Network, Data Mining techniques. Accuracy is 87%, Precision is 98% , Recall is 75%. In this particular paper the data set is collected and pre- processed using prestd function.While separating the dataset most of the data goes for as training set and very smaller portion will be considered as test data but in this particular paper 80% of data is taken as training data and remaining 20% of data is taken as test data set.

[26] This paper uses Seasonal trend decomposition using Loess (STL),Support vector machine(SVM),Random Forest,Back Propagation neural Network. Performanc-e is measured by comparing SVM and Random forest where SVM shows high performance than Random Forest. Dataset near Wuhan university is taken and computed using star topology network. With the help of this they can predict if rainfalls in particular area ,with that knowledge students can be given information to carry the umbrella.

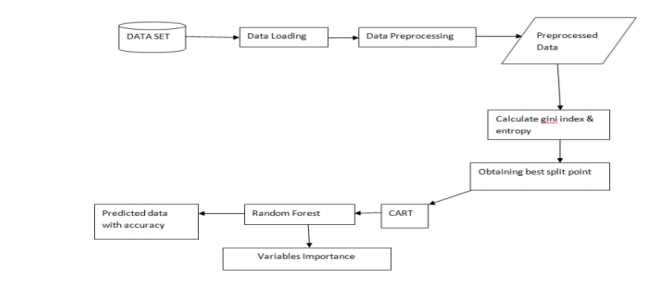
[27] A deep architecture combining the use of an auto encoder and a multilayer perceptron is used. Mean Square Error (MSE) and the Root Mean Square Error (RMSE). The data is extracted and divided into training dataset and testing dataset. Later it undergoes into normalisation. Vi = ai − minai maxai – minai. Where ai represents a value to normalize for the i-th variable, min ai is the minimum valor registered for this variable in the training set and maxai is the maximum valor registered for this variable in the training set. Dataset is divided into training, validation and testing.

[28] This paper discusses about Neural network, Classification and regression tree, Random forest, SVM, Boosted tree algorithm. Data is collected and pre processed. In this process rador images have been collected out of which it has a value -99 where they were not receiving any signals. and few null values they are removed. Different data mining algorithms have been used, out of SVM has been chosen as best model. For each model prediction metrics is done.

[29] This paper discusses about SVM, K nearest neighbours algorithms, variant K nearest neighbours ,AWS. In this they have generated new population vector from the original population a mutant vector is obtained by simply selecting two random vectors. No the mutant vector is crossed with the original vector and rest are called as trial vectors.

[30] This paper discusses about Adaptive neuro fuzzy inference system, Back propagation method, Hybrid method, Genetic algorithm, FIS. Since the forecasting of data is nonlinear systems ANFIS model is used in order to predict the previous year data. In this we have mean temperature,wind speed, relative humidity. The data is divided into 70% to 30%. 70% of data is for training period and 30% for validation period to develop ANFIS model.

**II. Flow Chart With Explanation:**



**Flowchat for the system**

**EXPLANATION:**

The datasets are preprocessed. It is fed as inputs for training. The rainfall values are clustered using subtractive clustering and the rainfall states identified as low, medium, heavy and given as outputs for training. Separating data into training and testing sets is an important part of evaluating data mining models. When we separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing. Here 80% of the dataset is used for training and the remaining 20 % for testing.

We will apply decision tree algorithm for it. In that Decision tree algorithm, by using Gini Index entropy method we will process the rainfall prediction.

**DATA COLLECTION AND DATA PREPROCESSING:**

Datasets for rainfall prediction downloaded from climatology information services (Hong Kong Observatory) and outliers and missing values are filtered using data cleaning process. In Data pre-processing, data cleaning, data integration, data transformation, data reduction takes place.

**GINI INDEX CALCULATION**:

We plot histogram for clear classification of each attribute based on its frequency and density to this processed dataset. Then, Gini index and entropy are applied to each and every attribute. Next, we plot Lc for each of these Gini and Entropy values. All the values of gini are stored in one variable and entropy is stored in another variable then graph is plotted for these two variables thereby showing the accuracy. When the data is sorted along with its corresponding class labels.

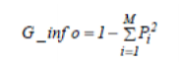
**FORMULAE:**

Evaluate the interval range [Interval range = (max – min)/group size]

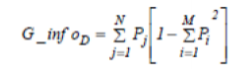
Based on this interval range, evaluate the split points whenever there is a change in the class label.

[split point= midpoint(changed class labels)]

Calculate the G\_info value for the class label by using



Calculate the G\_infoD value for each and every attribute by using

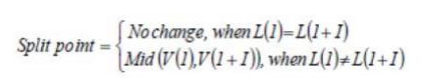


The Gini value is obtained by finding the difference between G\_info and G\_info D

values by using



The maximum gini value is considered the best split point and is the root node, as [Best Split Point = Maximum (Gini Index)] Repeat this procedure until every node ends with a unique class label. Now split points are found for particular set of intervals and group size in the processed data using



Apply the gini index value for each and every split point, compare all the gini index values from the first iteration and find the maximum value and its split value becomes the root node. From these, we get the decision tree based the classified rules.

**III Proposed Algorithm**

**DECISION TREE ALGORITHM:**

Decision tree algorithm using Gini Index in order to predict the precipitation with an accuracy and is completely based on the historical data. The decision tree is constructed and the classification rules are generated. To improve accuracy random forest technique is applied to this result thereby obtaining a result with increased accuracy rate. CART algorithm is also used for building decision tree. The dataset is divided into training and testing samples where we apply packages of party and r plot for training sample. Then, we test this system using the testing sample. From this, the misclassification rate is obtained**.**

**CONSTRUCTING DECISION TREE: (Step by step)**

**Step 1-** Take random samples and construct decision trees.

**Step 2-** Determine the importance of each attribute in the dataset.

**Step 3-** According to the weight, choose the best tree.

**Step 4-** According to the project perspective, the attribute that has higher importance is taken and the bar graph is plotted.

**Step 5-** According to the plotted graph, the value of highest frequency is taken as reference for accurate prediction.

**Step 6-** The dataset is divided into training and testing samples where we apply package (random forest) and

**Step 7-** Test this system using the testing sample.

**IV. Code**

library(readxl)

raindata <- read.csv("D:/deleted/3-2/Data Mining Techniques/project/raindata.csv")

View(raindata)

s = sum(is.na(raindata))

raindata <- na.omit(raindata)

x<-raindata[order(raindata$humidity),c(7,9)]

x

y<-raindata[order(raindata$temparature),c(4,9)]

y

z<-raindata[order(raindata$pressure),c(2,9)]

z

p<-raindata[order(raindata$windspeed),c(12,9)]

p

q<-raindata[order(raindata$dewpoint),c(6,9)]

q

r<-raindata[order(raindata$rainfall),c(9)]

r

hist(raindata$humidity,main="histogram of humidity",

xlab = "Humidity",

border="blue",col = "green",

xlim = c(36,98),las=0,

breaks = 10)

hist(raindata$temparature,main="histogram of temparature",

xlab = "temperature",

border="blue",col = "green",

freq = FALSE,

xlim = c(4,33),

las=1,

breaks = 10)

hist(raindata$pressure,main="histogram of pressure",

xlab = "pressure",

border="blue",col = "green",

xlim = c(998,1035),

las=1,

breaks = 10)

hist(raindata$windspeed,main="histogram of windspeed",

xlab = "windspeed",

border="blue",col = "green",

xlim = c(5,60),

las=1,

breaks = 10)

hist(raindata$winddirection,main="histogram of winddirection",

xlab = "winddirection",

border="blue",col = "green",

xlim = c(10,350),

las=1,

breaks = 10)

library(ineq)

g<-ineq(raindata$humidity,type="Gini")

g

# The Gini coefficient can then be thought of as the ratio of the area that lies

between the line of equality and the Lorenz curve (marked A in the diagram) over

the total area under the line of equality (marked A and B in the diagram);

plot(Lc(raindata$humidity),col="purple",lwd=2)#gini index

h<-ineq(raindata$temparature,type="Gini")

h

plot(Lc(raindata$temparature),col="green",lwd=2)#gini index

i<-ineq(raindata$dewpoint,type="Gini")

i

plot(Lc(raindata$dewpoint),col="blue",lwd=2)#gini index

j<-ineq(raindata$windspeed,type="Gini")

j

plot(Lc(raindata$windspeed),col="pink",lwd=2)#gini index

k<-ineq(raindata$pressure,type="Gini")

k

plot(Lc(raindata$pressure),col="orange",lwd=2)#gini index

a<-ineq(raindata$humidity,type="entropy")

a

plot(Lc(raindata$humidity),col="purple",lwd=2)

b<-ineq(raindata$temparature,type="entropy")

b

plot(Lc(raindata$temparature),col="green",lwd=2)

c<-ineq(raindata$dewpoint,type="entropy")

c<-0.06

plot(Lc(raindata$dewpoint),col="blue",lwd=2)

d<-ineq(raindata$windspeed,type="entropy")

d

plot(Lc(raindata$windspeed),col="pink",lwd=2)

e<-ineq(raindata$pressure,type="entropy")

e

plot(Lc(raindata$pressure),col="orange",lwd=2)

x1<-c(g,h,i,j,k)

x2<-c(a,b,c,d,e)

plot(x1,type="o",col="blue",ylim=c(0.0,0.4))

par(new=TRUE)

lines(x2,type="o",col="red")

data<-raindata

str(data)

data$rainfallf<-factor(data$rainfall)

#To ensure all results are reproducable

set.seed(12)

pd<-sample(2,nrow(data),replace=TRUE,prob=c(0.8,0.2))

train<-data[pd==1,]

validate<-data[pd==2,]

library(partykit)

library(tree)

library(party)

tree1<-ctree(rainfall~humidity+dewpoint+windspeed+sunshine,data=train)

tree1

#Conditional inference tree

plot(tree1,main="Conditional Inference Tree")

#table of prediction errors

tabl<-table(predict(tree1), train$rainfallf)

ctr<-sum(diag(tabl))/sum(tabl)

ctr

# Estimated class probabilities

predict(tree1,validate,type="prob")

predict(tree1,validate)

raindata$rainfall <- ifelse(raindata$cloud < 88, 'no', 'yes')

raindata$rainfall <- as.factor(raindata$rainfall)

table(raindata$rainfall)

set.seed(123)

samp <- sample(nrow(raindata), 0.6 \* nrow(raindata))

train <- raindata[samp, ]

test <- raindata[-samp, ]

library(randomForest)

model <- randomForest(rainfall ~

sunshine+temparature+windspeed+pressure+dewpoint+humidity -

cloud, data =train)

fit.rf <- randomForest(rainfall ~

sunshine+temparature+windspeed+pressure+dewpoint+humidity -

cloud, data = train)

print(fit.rf)

importance(fit.rf)

#plot(fit.rf)

#plot( importance(fit.rf), lty=2, pch=16)

#lines(importance(fit.rf))

varImpPlot(fit.rf,type=2)

imp = importance(fit.rf)

#take cloud as ref

pred <- predict(model, newdata = test)

pred

tab=table(pred, test$rainfall)

tab

rnd<-sum(diag(tab))/sum(tab)

rnd

ctr

rnd

**DATASET DESCRIPTION AND SAMPLE DATA:**

Datasets for rainfall prediction downloaded from climatology information services (Hong Kong Observatory) and outliers and missing values are filtered using data cleaning process. In Data preprocessing data cleaning, data integration, data transformation, data reduction takes place.

The datasets are preprocessed. It is fed as inputs for training. The rainfall values are clustered using subtractive clustering and the rainfall states identified as low, medium, heavy and given as outputs for training. Separating data into training and testing sets is an important part of evaluating data mining models. When we separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing. Here 80% of the dataset is used for training and the remaining 20 % for testing.



**REFERENCES**

[1] Valmik B Nikam , B.B.Meshram,”Modelling rainfall prediction using data mining method”, International Journal of Advanced Research in Computer Science and Software Engineering -Research Paper -Volume 5, Issue 10, October-2015

[2] Mugume , Waiswa D1, Mesquita MDS2, Reuder , Basalirwa Bamutaze, Twinomuhangi, Tumwine, Sansa Otim, Jacob Ngailo and Ayesiga,”Assessing the Performance of WRF Model in Simulating Rainfall over Western Uganda, Journal of Climatology & Weather Forecasting Mugume et al., J Climatol Weather Forecasting 2017.

[3] Narasimha Prasad , Krishna Rao Patro , Mannava Munirathnam Naidu,”A Gini Index BasedElegant Decision Tree Classifier to Predict Precipitation”, International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 .Volume 4 Issue 7, July 2015

[4] A.Geetha, Dr. G.M.Nasira, “Data mining for meteorological applications: Decision trees for modeling rainfall prediction”, International Journal Of Engineering And Computer Science ISSN:2319- 7242 Volume 3 Issue 5 may, 2016.

[5] Mr.. Dhawal hirani, Dr. Nitin Mishra,” A Survey On Rainfall Prediction Techniques”, International Journal of Computer Application (2250-1797) Volume 6– No.2, March- April 2016

[6] R.Senthil Kumar,Dr. C.Ramesh,”A Study on Prediction of Rainfall Using Datamining Technique”, International Journal of Emerging Technology and Advanced Engineering .ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 4, April 2014