

**School of Information Technology & Engineering**

**SWE 2009 – Data Mining Techniques**

**SLOT: A1+TA1**

**J Component**

**Review 1**

**Topic: RAINFALL PREDICTION USING DATA MINING**

**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**

**Objectives:**

To be able to predict the changes in the atmosphere for a particular location using Data mining techniques. Rainfall prediction poses right from the ancient times as a big herculean task, because it depends on various parameters to predict. Rainfall is one form of precipitation, and it primarily depends on humidity, temperature, pressure, wind speed, dew point, and so on. The present research is focused on using the gini index as an attribute selection measure in an elegant decision tree to predict precipitation for datasets, making data preprocessing and data transformation on raw weather data set the data mining, prediction model used for rainfall prediction. 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.

**Abstract:**

Rainfall prediction is an application which predicts the changes in the atmosphere for a location. We as humans are bound to make mistakes while predicting weather conditions which might result in damage to both life and property. To avoid this, we use data mining algorithms for early warning of climatic conditions such as like maximum temperature, minimum temperature wind speed, rainfall, humidity, pressure, dew point, cloud, sunshine and wind direction from data to predict rainfall.

**Literature Survey:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Sno** | **Title of** | **Algorithm s** | **Data set being** | **Performa** | **Gap** | **Scope for** |
|  | **the** | **used** | **used** | **nce** | **identified** | **future** |
|  | **Paper** |  |  | **measures** |  | **work** |
|  | **And** |  |  |  |  |  |
|  | **year** |  |  |  |  |  |
| 1 | Modelin | Five data- | Datasets are | Three | The | The |
|  | g and | mining | gotten from | metrics, | constructi | proposed |
|  | Predictio | algorithms, | radar data | the mean | on of a | methodolo |
|  | n of | neural | covering | absolute | predictive | gy has |
|  | Rainfall | network, | Oxford, radar | error | model is | demonstra |
|  | Using | random | and TB data | (MAE), | preceded | ted high- |
|  | Radar | forest, | collected at | mean | by | accuracy |
|  | Reflectiv | classification | Oxford, South | square | variable | rainfall |
|  | ity Data: | and | Amana (16 km | error | (input) | prediction |
|  | A Data- | regression | west of Oxford) | (MSE), | selection | s in |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Mining | tree, support | and Iowa City | and | aimed at | Oxford, |
|  | Approac | vector | (25 km east of | standard | the | Iowa. It |
|  | h | machine, and | Oxford) | deviation | reduction | has |
|  | 2013 | k-nearest |  | (SD), are | of the | provided |
|  |  | neighbor, are |  | used as the | dimension | useful |
|  |  | employed to |  | metrics for | ality and | hydrologic |
|  |  | build |  | prediction | complexit | al |
|  |  | prediction |  | accuracy | y of the | informatio |
|  |  | models. The |  |  | data set, as | n for |
|  |  | algorithm |  |  | well as | SWAT |
|  |  | offering the |  |  | gains in | models |
|  |  | highest |  |  | model | predicting |
|  |  | accuracy is |  |  | accuracy | water |
|  |  | selected for |  |  | and | quality |
|  |  | further study. |  |  | computati | [4]. In |
|  |  |  |  |  | onal | future |
|  |  |  |  |  | efficiency | research, |
|  |  |  |  |  |  | data from |
|  |  |  |  |  |  | other |
|  |  |  |  |  |  | regions |
|  |  |  |  |  |  | will be |
|  |  |  |  |  |  | collected |
|  |  |  |  |  |  | for further |
|  |  |  |  |  |  | validation |
|  |  |  |  |  |  | and |
|  |  |  |  |  |  | improvem |
|  |  |  |  |  |  | ent of the |
|  |  |  |  |  |  | proposed |
|  |  |  |  |  |  | approach |
| 2 | Compara | Forecasting | The rainfall | The model | The | It is |
|  | tive | techniques | datasets used in | compariso | analysis of | therefore |
|  | analysis | such as | this research | n is based | the models | recommen |
|  | of | Artificial | work were | on four | accuracy, | ded that |
|  | rainfall | Neural | collected from | criteria; | shows | the ANN |
|  | predictio | Network | an automatic | the Root | that, | and FL |
|  | n models | (ANN) and | weather station | Mean | overall, | techniques |
|  | using | Fuzzy Logic | in Iju, a town in | Square | the ANN | could be |
|  | neural | (FL) have | Akure North | Error | model | improved |
|  | network | been used to | Local | (RMSE), | perform | upon by |
|  | and | study rainfall. | Government | Mean | slightly | combining |
|  | fuzzy | The model | Area of Ondo | Absolute | better than | it with |
|  | logic | comparison is | State for the | Error | the FL | another |
|  | 2016 | based on four | period of four | (MAE), | model in | method |
|  |  | criteria; the | years (2007- | prediction | terms of | i.e., |
|  |  | Root Mean | 2010). | error, and | PE, | genetic |
|  |  | Square Error |  | the | RMSE, | algorithm |
|  |  | (RMSE), |  | prediction | MAE | for its |
|  |  | Mean |  | accuracy. | And | optimizati |
|  |  | Absolute |  |  | accuracy. | on |
|  |  | Error (MAE), |  |  |  | purpose. |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | prediction |  |  | The results |  |
|  |  | error, and the |  |  | show that |  |
|  |  | prediction |  |  | the neural |  |
|  |  | accuracy. |  |  | network |  |
|  |  |  |  |  | model is |  |
|  |  |  |  |  | better than |  |
|  |  |  |  |  | the fuzzy |  |
|  |  |  |  |  | logic |  |
|  |  |  |  |  | model. |  |
| 3 | An | This paper | A real world | The | The model | Although |
|  | artificial | presents a | case study was | common | performan | the model |
|  | neural | new approach | set up in | method | ce of 6 | performan |
|  | network | using an | Bangkok; 4 | used is | hour | ce of 6 |
|  | model | Artificial | years of hourly | trial and | forecastin | hour |
|  | for | Neural | data from 75 | error | g was low | forecastin |
|  | rainfall | Network | rain gauge | based on a | and the | g was low |
|  | forecasti | technique to | stations in the | total error | forecastin | and the |
|  | ng in | improve | area were used | criterion. | g was not | forecastin |
|  | Bangkok | rainfall | to develop the | This | as accurate | g was not |
|  | , | forecast | ANN model. | method | as | as |
|  | Thailand | performance. | The developed | starts with | expected, | accurate |
|  | 2009 |  | ANN model is | a small | the | as |
|  |  |  | being applied | number of | developed | expected, |
|  |  |  | for real time | nodes, | model can | the |
|  |  |  | rainfall | gradually | still be | developed |
|  |  |  | forecasting and | increasing | used for | model can |
|  |  |  | flood | the | practical | still be |
|  |  |  | management in | network | applicatio | used for |
|  |  |  | Bangkok, | size until | ns such as | practical |
|  |  |  | Thailand | the desired | rainfall | applicatio |
|  |  |  |  | accuracy | forecastin | ns such as |
|  |  |  |  | is | g and | rainfall |
|  |  |  |  | achieved. | flood | forecastin |
|  |  |  |  |  | manageme | g and |
|  |  |  |  |  | nt for the | flood |
|  |  |  |  |  | urban | manageme |
|  |  |  |  |  | areas. | nt for the |
|  |  |  |  |  |  | urban |
|  |  |  |  |  |  | areas. |
| 4 | Modelin | They have | Weather data is | The model | When a | This |
|  | g | used | one of the | has | predictor | model is |
|  | Rainfall | Bayesian | meteorological | simplicity, | category is | nearly |
|  | Predictio | classification | data that is rich | good | not present | accurate |
|  | n Using | model on | with important | prediction | in the | model in |
|  | Data | different | information, | performan | training | compariso |
|  | Mining | datasets from | which can be | ce, and | data, the | n with |
|  | Method: | various | used for weather | can be | model | compute |
|  | A | places. The | prediction they | used for | assumes | intensive |
|  | Bayesian | model can be | extracted | both | that a new | models. |
|  |  | deployed on | weather data | binary and | record | The |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Approac | commodity | historical data | multiclass | with that | performan |
|  | h | hardware; do | collected from | prediction | category | ce of the |
|  | 2013 | not demand | Indian | problems | has zero | model can |
|  |  | high- | Meteorological |  | probability | also be |
|  |  | performance | Department |  | . This | improved |
|  |  | cluster or | (IMD) Pune. |  | could be a | by |
|  |  | supercomputi | From the |  | major | designing |
|  |  | ng | collected |  | issue if | the model |
|  |  | environment. | weather data |  | this rare | for |
|  |  |  | comprising of |  | predictor | scalable |
|  |  |  | 36 attributes, |  | value is | platforms, |
|  |  |  | only 7 attributes |  | important. | either for |
|  |  |  | are most |  |  | vertical |
|  |  |  | relevant to |  |  | scalability |
|  |  |  | rainfall |  |  | or for |
|  |  |  | prediction |  |  | horizontal |
|  |  |  |  |  |  | scalability. |
| 5 | A Study | In this | Monthly | The | Regressio | In the |
|  | on | research | rainfall(in MM) | parameters | n | future |
|  | Predictio | paper, | data for Coastal | include the | technique | works, |
|  | n of | analysis of | Andhra, | Root | could not | some |
|  | Rainfall | various | Telangana and | Mean | find | additional |
|  | Using | popular data | Rayalaseema | Square | accurate | inputs |
|  | Data | mining | regions in | Error | value of | were |
|  | mining | algorithms is | Andhra Pradesh | RMSE, | prediction. | employed |
|  | Techniqu | presented for | state during the | Mean | May work | for rainfall |
|  | e 2016 | rainfall | years 1871-2011 | Absolute | poor for | prediction |
|  |  | prediction. | is collected from | Error | small data | such as |
|  |  | Recent | Climatology & | MAE, | sets. | Sea |
|  |  | algorithms | Hydrometeorolo | Coefficien |  | Surface |
|  |  | analyzed in | gy Division, | t Of |  | Temperatu |
|  |  | this work are | Indian Institute | Correlatio |  | re (SST) |
|  |  | Naive Bayes, | of Tropical | n CC and |  | areas |
|  |  | K- Nearest | Meteorology | BIAS. |  | around |
|  |  | Neighbour | (IITM), Pune, | Two-Third |  | Andhra |
|  |  | algorithm, | India. | of the data |  | Pradesh |
|  |  | Decision |  | was used |  | and |
|  |  | Tree, ANFIS, |  | for |  | Southern |
|  |  | ARIMA, |  | training |  | part of |
|  |  | SLIQ, Neural |  | the model |  | India. |
|  |  | Network and |  | and One- |  |  |
|  |  | Fuzzy logic |  | third for |  |  |
|  |  | are some of |  | testing |  |  |
|  |  | the |  |  |  |  |
|  |  | algorithms |  |  |  |  |
|  |  | compared in |  |  |  |  |
|  |  | this paper. |  |  |  |  |
| 6 | A Gini | This research |  | The Root | The main | As |
|  | Index | is focused on | Four data sets | Mean | advantage | precipitati |
|  | Based | using the | have used, out | Square | of this | on |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Elegant | Gini index as | of which three | Error | algorithm, | depends |
|  | Decision | an attribute | datasets are of | (RMSE), | when | on |
|  | Tree | selection | actual cities | Mean | compared | dynamic |
|  | Classifie | measure in an | data. We have | Absolute | to SLIQ, | attributes, |
|  | r to | elegant | used monsoon | Error | is the | a decision |
|  | Predict | decision tree | period data | (MAE), | reduction | tree must |
|  | Precipita | to predict |  | prediction | in | be |
|  | tion | precipitation |  | error | computati | developed |
|  | 2013 | for |  |  | onal | for the |
|  |  | voluminous |  |  | complexit | dynamic |
|  |  | datasets. |  |  | y. Overall, | data mode |
|  |  | A dynamic |  |  | this | rather than |
|  |  | model has |  |  | reduction | static data |
|  |  | been |  |  | eventually | mode. |
|  |  | developed by |  |  | results in | Here, can |
|  |  | using image |  |  | less time | improve |
|  |  | processing |  |  | for | by using |
|  |  | with data |  |  | execution | dynamic |
|  |  | mining |  |  | and better | attributes. |
|  |  | techniques to |  |  | classificati |  |
|  |  | analyse |  |  | on |  |
|  |  | satellite |  |  | accuracy |  |
|  |  | images for |  |  | than the |  |
|  |  | the prediction |  |  | SLIQ and |  |
|  |  | of |  |  | neural |  |
|  |  | precipitation. |  |  | network |  |
|  |  |  |  |  | techniques |  |
|  |  |  |  |  | . |  |
|  |  |  |  |  |  |  |
| 7 | Assessin | The root | In the study, | The root | Results | The GF |
|  | g the | mean square | daily observed | mean | show that | cumulus |
|  | Performa | square error, | rainfall data | square | the Grell- | scheme as |
|  | nce of | mean error | from 21st April | error, | Fretas | a basic |
|  | WRF | and the sign | to 10th May | mean error | scheme is | determinis |
|  | Model in | test method | 2013 for five | and the | better at | tic scheme |
|  | Simulati | are used to | stations of | sign test | simulating | which can |
|  | ng | assess the | western Uganda | method | rainfall | also be |
|  | Rainfall | ability of the | (Table 1) was | are used to | compared | used as a |
|  | over | schemes to | obtained from | assess the | to other | convection |
|  | Western | simulate | Uganda | ability of | schemes | scheme to |
|  | Uganda | rainfall along | National | the | over the | give a |
|  | 2017 | with an | Meteorological | schemes to | study | backgroun |
|  |  | adapted | Authority | simulate | period | d analysis |
|  |  | contingency | (UNMA) | rainfall | while the | during |
|  |  | table. |  | along with | Betts- | data |
|  |  |  |  | an adapted | Miller- | assimilatio |
|  |  |  |  | contingenc | Janji’c and | n but for |
|  |  |  |  | y table. | the Kain- | heavy |
|  |  |  |  |  | Fritsch | rainfall |
|  |  |  |  |  | schemes | events, |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | overestima | BMJ |
|  |  |  |  |  | ted | scheme is |
|  |  |  |  |  | rainfall. | recommen |
|  |  |  |  |  |  | ded |
|  |  |  |  |  |  |  |
| 8 | Data | They | This paper | The | Though | This |
|  | mining | mentioned | highlights a | proposed | the | model |
|  | for | notable | model using | model is | performan | may be |
|  | meteorol | decision-tree | decision tree to | implement | ce vector | used in |
|  | ogical | algorithms | predict weather | ed using | depicts | machine |
|  | applicati | like ID3 | phenomena like | the open | accuracy | learning |
|  | ons: | (Iterative | fog, rainfall, | source | as 100%, | and |
|  | Decision | dichotomiser | cyclones and | data | when the | further |
|  | trees for | 3), C4.5 | thunderstorms, | mining | compariso | promises |
|  | modeling | (successor of | which can be a | tool | n is done | the scope |
|  | rainfall | ID3), CART | life saving | Rapidmine | between | for |
|  | predictio | (Classificatio | information and | r for | the actual | improvem |
|  | n | n And | used by peoples | accurate | and target | ent as |
|  | 2015 | Regression | of all walks of | measurem | data, we | more and |
|  |  | Tree), | life in making | ents. | find that | more |
|  |  | CHAID | wise and |  | the | relevant |
|  |  | (CHI-squared | intelligent |  | success | attributes |
|  |  | Automatic | decisions. |  | rate for the | can be |
|  |  | Interaction |  |  | year 2014 | used in |
|  |  | Detector) and |  |  | is 80.67. | predicting |
|  |  | MARS |  |  |  | the |
|  |  | (extends |  |  |  | dependent |
|  |  | decision trees |  |  |  | variables. |
|  |  | to better |  |  |  |  |
|  |  | handle |  |  |  |  |
|  |  | numerical |  |  |  |  |
|  |  | data.) |  |  |  |  |
| 9 | A Survey | Widely used | The Empirical | Multiple | One of the | some |
|  | on | techniques | approach is | Linear | challenges | limitations |
|  | Rainfall | for prediction | based on | Regressio | that face | is clearly |
|  | Predictio | are | analysis of past | n, | the | noticed in |
|  | n | Regression | historical data of | Autoregre | knowledge | all the |
|  | Techniqu | analysis, | weather and its | ssive | discovery | methods |
|  | es | clustering, | relationship to a | Integrated | process in | of rainfall |
|  | 2016 | and Artificial | variety of | Moving | meteorolo | prediction |
|  |  | Neural | atmospheric | Average | gical data | discussed |
|  |  | Network | variables over | (ARIMA) | is poor | in this |
|  |  | (ANN) etc. | different parts of | Model, | data | survey |
|  |  |  | Chhattisgarh. | Genetic | quality. | paper The |
|  |  |  |  | algorithm, | For this | extensive |
|  |  |  |  | Adaptive | reason | references |
|  |  |  |  | Splines | they try to | in support |
|  |  |  |  | Threshold | prepare | of the |
|  |  |  |  | Autoregre | their data | different |
|  |  |  |  | ssive | carefully | developme |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | (ASTAR) | to obtain | nts of |
|  |  |  |  | Modelling | accurate | methods |
|  |  |  |  | is used for | and | provided |
|  |  |  |  | accurate | correct | in this |
|  |  |  |  | measurem | results. | research |
|  |  |  |  | ents | First they | should be |
|  |  |  |  |  | choose the | of great |
|  |  |  |  |  | most | help to |
|  |  |  |  |  | related | researcher |
|  |  |  |  |  | attributes | s to |
|  |  |  |  |  | to their | accurately |
|  |  |  |  |  | mining | predict |
|  |  |  |  |  | task. | rainfall in |
|  |  |  |  |  | purpose | the future |
|  |  |  |  |  | they | and to |
|  |  |  |  |  | neglect the | select the |
|  |  |  |  |  | wind | method |
|  |  |  |  |  | direction. | that would |
|  |  |  |  |  | Then they | solve their |
|  |  |  |  |  | remove | problem |
|  |  |  |  |  | the | they will |
|  |  |  |  |  | missing | be facing |
|  |  |  |  |  | value | in their |
|  |  |  |  |  | records | proposed |
|  |  |  |  |  |  | prediction |
|  |  |  |  |  |  | model. |
| 10 | Feature | Differential | The weather | The k-NN | Evolutiona | In future |
|  | Selection | Evolutionary | data, which are | algorithm | ry | work, they |
|  | for Very | algorithms | collected from | is amongst | computati | will |
|  | Short- | were used to | 408 automatic | the | on does | preprocess |
|  | Term | select | weather stations | simplest of | not use | the |
|  | Heavy | important | during the recent | all | binary | weather |
|  | Rainfall | features. | four years from | machine | encoding | data by |
|  | Predictio | Discriminant | 2007 to 2010, | learning | as a | various |
|  | n Using | functions, | had a | algorithms | simple | methods, |
|  | Evolutio | such as | considerable | . | genetic | such as |
|  | nary | support | number of | selected | algorithm | representat |
|  | Computa | vector | missing data, | features | The | ion |
|  | tion | machine | erroneous data, | were used | weather | learning, |
|  | 2014 | (SVM), k- | and unrelated | to predict | data | cyclic |
|  |  | nearest | features. They | very short- | collected | loess, |
|  |  | neighbors | analyzed the | term | from | contrast, |
|  |  | algorithm (k- | data and | heavy | automatic | and |
|  |  | NN),Normali | corrected the | rainfall. | weather | quantile |
|  |  | zation. | errors. They | The results | stations | normalizat |
|  |  |  | preprocessed the | of GA | during the | ion |
|  |  |  | original data | were | recent four | algorithms |
|  |  |  | given by KMA | significant | years had | . Also, |
|  |  |  |  | ly better | a lot of | other |
|  |  |  |  |  | missing | machine |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | than those | data and | learning |
|  |  |  |  | of DE. | erroneous | techniques |
|  |  |  |  |  | data. | , such as |
|  |  |  |  |  |  | statistical |
|  |  |  |  |  |  | relational |
|  |  |  |  |  |  | learning, |
|  |  |  |  |  |  | multilinear |
|  |  |  |  |  |  | subspace |
|  |  |  |  |  |  | learning, |
|  |  |  |  |  |  | and |
|  |  |  |  |  |  | associatio |
|  |  |  |  |  |  | n rule |
|  |  |  |  |  |  | learning |
|  |  |  |  |  |  | can be |
|  |  |  |  |  |  | applied in |
|  |  |  |  |  |  | future. |
| 11 | BPN | Back | rainfall data of | composite | These two | The main |
|  | Based | Propagation | past years to | network is | techniques | advantage |
|  | Rainfall | Network | forecast future | much | fail to | of the |
|  | Forecasti | (BPN) | rainfall | more | detect new | BPN |
|  | ng: A |  |  | efficient | attacks | neural |
|  | Review |  |  | than a | that are | network |
|  | 2017 |  |  | conventio | not present | method is |
|  |  |  |  | nal neural | in the | that it can |
|  |  |  |  | network. | training | fairly |
|  |  |  |  |  | data set. | approxima |
|  |  |  |  |  | We | te a large |
|  |  |  |  |  | improve | class of |
|  |  |  |  |  | them for | functions. |
|  |  |  |  |  | anomaly |  |
|  |  |  |  |  | intrusion |  |
|  |  |  |  |  | detection |  |
|  |  |  |  |  | and test |  |
|  |  |  |  |  | them over |  |
|  |  |  |  |  | the KDD |  |
|  |  |  |  |  | 99 data |  |
|  |  |  |  |  | sets and |  |
|  |  |  |  |  | over real |  |
|  |  |  |  |  | network |  |
|  |  |  |  |  | traffic in |  |
|  |  |  |  |  | real time. |  |
| 12 | RAINFA | linear | Monthly | percentage | If we want | In the |
|  | LL | techniques, | rainfall(in MM) | better | to predict | future |
|  | PREDIC | such as auto | data for Coastal | compariso | numbers | works, |
|  | TION | regressive | Andhra, | n with | before | some |
|  | USING | (AR),Autoreg | Telangana and | SARIMA | they | additional |
|  | DATA | ressive– | Rayalaseema | models. | occur, | inputs |
|  | MINING | moving- | regions in |  | then | were |
|  | TECHNI | average | Andhra Pradesh |  | regression | employed |
|  | QUES - | model with | state during the |  | methods | for rainfall |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | A | exogenous | years 1871-2011 |  | are used. | prediction |
|  | SURVE | inputs | is collected from |  | Linear | such as |
|  | Y | (ARMAX), | Climatology & |  | regression | Sea |
|  | 2010 | and | Hydrometeorolo |  | is one of | Surface |
|  |  | KalmanFilter | gy Division, |  | the | Temperatu |
|  |  | (KF) | Indian Institute |  | regression | re (SST) |
|  |  |  | of Tropical |  | methods, | areas |
|  |  |  | Meteorology |  | and one of | around |
|  |  |  | (IITM), Pune, |  | the | Andhra |
|  |  |  | India |  | algorithms | Pradesh |
|  |  |  |  |  | tried out | and |
|  |  |  |  |  | first by | Southern |
|  |  |  |  |  | most | part of |
|  |  |  |  |  | machine | India. |
|  |  |  |  |  | learning |  |
|  |  |  |  |  | profession |  |
|  |  |  |  |  | als.If there |  |
|  |  |  |  |  | is a need |  |
|  |  |  |  |  | to classify |  |
|  |  |  |  |  | objects or |  |
|  |  |  |  |  | categories |  |
|  |  |  |  |  | based on |  |
|  |  |  |  |  | their |  |
|  |  |  |  |  | historical |  |
|  |  |  |  |  | classificati |  |
|  |  |  |  |  | ons and |  |
|  |  |  |  |  | attributes, |  |
|  |  |  |  |  | then |  |
|  |  |  |  |  | classificati |  |
|  |  |  |  |  | on |  |
|  |  |  |  |  | methods |  |
|  |  |  |  |  | like |  |
|  |  |  |  |  | decision |  |
|  |  |  |  |  | trees are |  |
|  |  |  |  |  | used. |  |
| 13 | Rainfall | The proposed | Daily rainfall | Regressio | From | The |
|  | Predictio | ENN used | datasets were | n | the explan | research |
|  | n using | Artificial | collected from | Analysis, | ation of | focus on |
|  | Data | Neural | 42 cities of two | Mean | how each | the |
|  | Mining | Networks and | continents, with | Square | model | domain of |
|  | Techniqu | Genetic | very diverse | Error and | handles | rainfall |
|  | es: A | Algorithm to | climatic | Magnitude | data, one | prediction |
|  | Systemat | identify the | features. | of Relative | would | has been |
|  | ic | best weights |  | Error | assume | increasing |
|  | Literatur | and biases. |  |  | that | since last |
|  | e Review |  |  |  | decision | decade |
|  | 5, 2018 |  |  |  | trees are a | and so are |
|  |  |  |  |  | better | the |
|  |  |  |  |  | model. In | problem |
|  |  |  |  |  | fact, I | areas. So |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | have seen | it was |
|  |  |  |  |  | the Neural | concluded |
|  |  |  |  |  | Network a | that |
|  |  |  |  |  | chieve 99 | enhancem |
|  |  |  |  |  | % | ents, |
|  |  |  |  |  | accuracy | optimizati |
|  |  |  |  |  | on a data | ons and |
|  |  |  |  |  | set while | integration |
|  |  |  |  |  | the | s of data |
|  |  |  |  |  | decision | mining |
|  |  |  |  |  | tree model | methods |
|  |  |  |  |  | only achie | are vital to |
|  |  |  |  |  | ved 86% | explore |
|  |  |  |  |  | accuracy | and solve |
|  |  |  |  |  | on the | these |
|  |  |  |  |  | same data | problems. |
|  |  |  |  |  | set. The |  |
|  |  |  |  |  | best fitted |  |
|  |  |  |  |  | model is |  |
|  |  |  |  |  | the one |  |
|  |  |  |  |  | that most |  |
|  |  |  |  |  | accurately |  |
|  |  |  |  |  | fits your |  |
|  |  |  |  |  | data |  |
| 14 | Rainfall | Artificial | three months | Measurem | From | Rainfall is |
|  | Forecasti | intelligence | rainfall data of | ent and | the explan | most |
|  | ng Using | and neural | particular region | years plot | ation of | essential |
|  | Data | network | for five years | in the | how each | for our |
|  | Mining |  |  | graph | model | life. So, |
|  | Techniqu |  |  | using x | handles | we predict |
|  | e |  |  | and y axis. | data, one | that |
|  | 2010, |  |  |  | would | rainfall in |
|  |  |  |  |  | assume | the certain |
|  |  |  |  |  | that | period. |
|  |  |  |  |  | decision | Therefore, |
|  |  |  |  |  | trees are a | we avoid |
|  |  |  |  |  | better | flood, |
|  |  |  |  |  | model. In | cyclone, |
|  |  |  |  |  | fact, I | forest fire |
|  |  |  |  |  | have seen | detection, |
|  |  |  |  |  | the Neural | global |
|  |  |  |  |  | Network a | warming |
|  |  |  |  |  | chieve 99 | etc. In |
|  |  |  |  |  | % | future we |
|  |  |  |  |  | accuracy | predict the |
|  |  |  |  |  | on a data | rainfall |
|  |  |  |  |  | set while | forecastin |
|  |  |  |  |  | the | g and |
|  |  |  |  |  | decision | other |
|  |  |  |  |  | tree model | applicatio |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | only achie | ns done by |
|  |  |  |  |  | ved 86% | using the |
|  |  |  |  |  | accuracy | artificial |
|  |  |  |  |  | on the | intelligenc |
|  |  |  |  |  | same data | e, neural |
|  |  |  |  |  | set. The | network |
|  |  |  |  |  | best fitted | and fuzzy |
|  |  |  |  |  | model is | sets etc. |
|  |  |  |  |  | the one | We do the |
|  |  |  |  |  | that most | research |
|  |  |  |  |  | accurately | on public |
|  |  |  |  |  | fits your | sectors |
|  |  |  |  |  | data | and save |
|  |  |  |  |  |  | the world. |
| 15 | Data | decision tree | The Rainfall | air | - | This |
|  | mining | algorithm | dataset is | temperatur |  | research |
|  | techniqu |  | collected from | e, wind |  | can be |
|  | es for |  | SPICE | data, soil |  | extended |
|  | rainfall |  | RESEARCH | temperatur |  | with more |
|  | predictio |  | CENTER – | e, surface |  | oceanic |
|  | n in the |  | TEPI (a | temperatur |  | pa- |
|  | Tepi |  | Government | e etc |  | rameters, |
|  | region of |  | office) |  |  | and other |
|  | Ethiopia |  |  |  |  | parameters |
|  | 2018 |  |  |  |  | like air |
|  |  |  |  |  |  | temperatur |
|  |  |  |  |  |  | e, surface |
|  |  |  |  |  |  | tem- |
|  |  |  |  |  |  | perature, |
|  |  |  |  |  |  | and soil |
|  |  |  |  |  |  | temperatur |
|  |  |  |  |  |  | e etc. The |
|  |  |  |  |  |  | other |
|  |  |  |  |  |  | aspect of |
|  |  |  |  |  |  | taking this |
|  |  |  |  |  |  | research |
|  |  |  |  |  |  | further is |
|  |  |  |  |  |  | by |
|  |  |  |  |  |  | comparing |
|  |  |  |  |  |  | the results |
|  |  |  |  |  |  | by |
|  |  |  |  |  |  | adopting |
|  |  |  |  |  |  | different |
|  |  |  |  |  |  | methodolo |
|  |  |  |  |  |  | gy and |
|  |  |  |  |  |  | algorithms |
|  |  |  |  |  |  | like |
|  |  |  |  |  |  | Clustering |
|  |  |  |  |  |  | , ANN and |
|  |  |  |  |  |  | Fuzzy |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  | logic. This |
|  |  |  |  |  |  | will give a |
|  |  |  |  |  |  | new vision |
|  |  |  |  |  |  | for |
|  |  |  |  |  |  | selecting |
|  |  |  |  |  |  | the |
|  |  |  |  |  |  | appropriat |
|  |  |  |  |  |  | e |
|  |  |  |  |  |  | algorithm |
|  |  |  |  |  |  | based on |
|  |  |  |  |  |  | the |
|  |  |  |  |  |  | different |
|  |  |  |  |  |  | parameters |
|  |  |  |  |  |  | that are |
|  |  |  |  |  |  | considered |
|  |  |  |  |  |  | for our |
|  |  |  |  |  |  | prediction |
|  |  |  |  |  |  | analysis. |
| 16 | Rainfall | Support | dataset is | in terms of | The | It is |
|  | Predictio | Vector | obtained from | precision, | biggest | suggested |
|  | n in | Machine, | weather | recall and | difference | for future |
|  | Lahore | Naïve Bayes, | forecasting | f measure, | between | work that |
|  | City | k Nearest | website from |  | the two | further |
|  | using | Neighbor, | December 1, |  | algorithms | prediction |
|  | Data | Decision | 2005 to |  | is that | s should |
|  | Mining | Tree and | November 31, |  | SVM uses | be |
|  | Techniqu | Multilayer | 2017 (12 years), |  | the kernel | performed |
|  | es | Perceptron | which contains |  | trick to | by |
|  | 2018 |  | several weather |  | turn a | exploring |
|  |  |  | related attributes |  | linearly | more |
|  |  |  | such as |  | nonsepara | classificati |
|  |  |  | Temperature, |  | ble | on |
|  |  |  | Atmospheric |  | problem | techniques |
|  |  |  | pressure, |  | into a | and |
|  |  |  | Relative |  | linearly | climatic |
|  |  |  | humidity etc |  | separable | attributes |
|  |  |  |  |  | one | on |
|  |  |  |  |  | (unless of | different |
|  |  |  |  |  | course we | weather |
|  |  |  |  |  | use the | data. |
|  |  |  |  |  | linear |  |
|  |  |  |  |  | kernel), |  |
|  |  |  |  |  | while |  |
|  |  |  |  |  | decision |  |
|  |  |  |  |  | trees (and |  |
|  |  |  |  |  | forests |  |
|  |  |  |  |  | based on |  |
|  |  |  |  |  | them, and |  |
|  |  |  |  |  | boosted |  |
|  |  |  |  |  | trees, both |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | to a lesser |  |
|  |  |  |  |  | extent due |  |
|  |  |  |  |  | to the |  |
|  |  |  |  |  | nature of |  |
|  |  |  |  |  | the |  |
|  |  |  |  |  | ensemble |  |
|  |  |  |  |  | algorithms |  |
|  |  |  |  |  | ) split the |  |
|  |  |  |  |  | input |  |
|  |  |  |  |  | space into |  |
|  |  |  |  |  | hyper- |  |
|  |  |  |  |  | rectangles |  |
|  |  |  |  |  | according |  |
|  |  |  |  |  | to the |  |
|  |  |  |  |  | target. |  |
| 17 | RAINFA | Decision | weather data |  | As the | Future |
|  | LL | Tree, | from January | True Dry | Bayesian | work can |
|  | PREDIC | Bayesian | 2011 to | True | network | include |
|  | TION | Technique. | December of | Rainy | provides a | expanded |
|  | USING |  | 2016 used as a | Class | global | database |
|  | DATA |  | dataset. | Precision | view of | with other |
|  | MINING |  |  |  | variables | important |
|  | TECHNI |  |  | Clas | associatio | weather |
|  | QUES |  |  | Recall | ns, the | parameters |
|  | 2010 |  |  | Pred.Dry | decision | and |
|  |  |  |  |  | tree is | include |
|  |  |  |  |  | more | using this |
|  |  |  |  |  | easily | weather |
|  |  |  |  |  | interpretab | informatio |
|  |  |  |  |  | le. These | n in |
|  |  |  |  |  | approache | agriculture |
|  |  |  |  |  | s provide | sector |
|  |  |  |  |  | insights on | reform |
|  |  |  |  |  | the current | with |
|  |  |  |  |  | care | cutting |
|  |  |  |  |  | process. | edge |
|  |  |  |  |  |  | technologi |
|  |  |  |  |  |  | es. |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 18 | PREDIC | K-clustering, | dataset of | Oceanic | If we have | Currently |
|  | TION | Decision tree | rainfall for the | Nino | no idea | we have |
|  | OF | and | past couple of | Index | about the | focused on |
|  | RAINFA | Regression | years | (ONI) is | data and | 34 stations |
|  | LL | model |  | the | want to | across |
|  | USING |  |  | standard | group data | Banglades |
|  | DATA |  |  |  | points to | h. We |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | MINING |  |  | measurem | understand | would like |
|  | TECHNI |  |  | ent | their | to increase |
|  | QUES |  |  |  | collective | the range |
|  | 2013 |  |  |  | behavior, | of our |
|  |  |  |  |  | clustering | capabilitie |
|  |  |  |  |  | is one of | s. Such |
|  |  |  |  |  | the go-to | initiation |
|  |  |  |  |  | methods. | will |
|  |  |  |  |  | If there is | enable |
|  |  |  |  |  | a need to | farmers to |
|  |  |  |  |  | classify | do |
|  |  |  |  |  | objects or | efficient |
|  |  |  |  |  | categories | farming |
|  |  |  |  |  | based on | across the |
|  |  |  |  |  | their | country. |
|  |  |  |  |  | historical | Increasing |
|  |  |  |  |  | classificati | the range |
|  |  |  |  |  | ons and | will also |
|  |  |  |  |  | attributes, | give rise |
|  |  |  |  |  | then | to the |
|  |  |  |  |  | classificati | number of |
|  |  |  |  |  | on | data in the |
|  |  |  |  |  | methods | system. |
|  |  |  |  |  | like |  |
|  |  |  |  |  | decision |  |
|  |  |  |  |  | trees are |  |
|  |  |  |  |  | used. |  |
| 19 | Study of | subtractive | minimum | stream | If we have | In this |
|  | Various | clustering | temperature, | flow | no idea | survey |
|  | Rainfall |  | maximum | interpolate | about the | paper we |
|  | Estimati |  | temperature | d rainfall | data and | found the |
|  | on & |  | ,pressure, wind |  | want to | use of |
|  | Predictio |  | direction |  | group data | various |
|  | n |  | ,relative |  | points to | data |
|  | Techniqu |  | humidity etc |  | understand | mining |
|  | es Using |  |  |  | their | techniques |
|  | Data |  |  |  | collective | on the |
|  | Mining |  |  |  | behavior, | collected |
|  | 2017 |  |  |  | clustering | data set |
|  |  |  |  |  | is one of | from the |
|  |  |  |  |  | the go-to | various |
|  |  |  |  |  | methods. | resources |
|  |  |  |  |  | If there is | may found |
|  |  |  |  |  | a need to | useful in |
|  |  |  |  |  | classify | accurate |
|  |  |  |  |  | objects or | prediction |
|  |  |  |  |  | categories | of |
|  |  |  |  |  | based on | rainfall.It |
|  |  |  |  |  | their | is |
|  |  |  |  |  | historical | observed |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | classificati | from |
|  |  |  |  |  | ons and | various |
|  |  |  |  |  | attributes, | studies |
|  |  |  |  |  | then | that |
|  |  |  |  |  | classificati | rainfall |
|  |  |  |  |  | on | estimation |
|  |  |  |  |  | methods | and |
|  |  |  |  |  | like | prediction |
|  |  |  |  |  | decision | varies |
|  |  |  |  |  | trees are | from using |
|  |  |  |  |  | used. | MLR to |
|  |  |  |  |  |  | SLIQ. |
| 20 | RAINFA | mathematical | The data set of | Average | Linear | We may |
|  | LL | method | 100 years is | Temperatu | regression | improve |
|  | PREDIC | called Linear | taken for this | re | is a linear | this |
|  | TION | Regression | project and the | Coefficien | model, | system |
|  | USING |  | implemented | t | which | further |
|  | MODIFI |  | using Numerical |  | means it | using |
|  | ED |  | methods | Cloud | works | multiple |
|  | LINEAR |  |  | Cover | really | regression |
|  | REGRE |  |  | Coefficien | nicely | which can |
|  | SSIO N |  |  | t | when the | take |
|  |  |  |  |  | data has a | multiple |
|  |  |  |  | Constant | linear | months at |
|  |  |  |  | Value | shape. | a time as |
|  |  |  |  |  | But, when | input and |
|  |  |  |  | Iterations | the data | just |
|  |  |  |  |  | has a non- | forming a |
|  |  |  |  | Times | linear | single |
|  |  |  |  | Performed | shape, | equation |
|  |  |  |  |  | then a | which |
|  |  |  |  |  | linear | leads |
|  |  |  |  |  | model | nearer to |
|  |  |  |  |  | cannot | an |
|  |  |  |  |  | capture the | accurate |
|  |  |  |  |  | non-linear | rainfall |
|  |  |  |  |  | features. | predicted. |
|  |  |  |  |  | So in this | The |
|  |  |  |  |  | case, you | proposed |
|  |  |  |  |  | can use | approach |
|  |  |  |  |  | the | may also |
|  |  |  |  |  | decision | be used in |
|  |  |  |  |  | trees, | other |
|  |  |  |  |  | which do a | applicatio |
|  |  |  |  |  | better job | ns like, in |
|  |  |  |  |  | at | schools to |
|  |  |  |  |  | capturing | predict the |
|  |  |  |  |  | the non- | average |
|  |  |  |  |  | linearity in | marks of |
|  |  |  |  |  | the data by | their |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | dividing | students, |
|  |  |  |  |  | the space | in sports |
|  |  |  |  |  | into | to predict |
|  |  |  |  |  | smaller | the scores |
|  |  |  |  |  | sub- | or winning |
|  |  |  |  |  | spaces. | teams |
|  |  |  |  |  |  | based on |
|  |  |  |  |  |  | their |
|  |  |  |  |  |  | previous |
|  |  |  |  |  |  | performan |
|  |  |  |  |  |  | ce, in |
|  |  |  |  |  |  | enterprises |
|  |  |  |  |  |  | to estimate |
|  |  |  |  |  |  | their |
|  |  |  |  |  |  | profits, |
|  |  |  |  |  |  | etc.,. |
| 21 | Predictio | 1)K means | National Centres | Since we | In this | This |
|  | n of | clustering | for | are using | particular | relationshi |
|  | daily | technique | Environmental | clustering | paper | p is then |
|  | rainfall | 2)Classificati | Prediction/Natio | techniques | CART | applied to |
|  | state in a | on and | nal Centre for | it require | model is | the |
|  | river | regression | Atmospheric | many | trained | General |
|  | basin | Tree | Research | clusters as | and a | Circulatio |
|  | using | 3)Statistical | (NCEP/NCAR) | input | relationshi | n Model |
|  | statistical | downscaling | reanalysis | which is | p is | which is - |
|  | downscal | method is | climatic data set | major | establishe | simulated, |
|  | ing from | used to |  | problem to | d between | standardiz |
|  | GCM | predict the |  | overcome | the states | ed, bias |
|  | output | future rainfall |  | we use | where the | free large- |
|  | 2011 | places using |  | validation | rainfall is | scale |
|  |  | GCM |  | measures | happen | climate |
|  |  | simulated |  | and they | near the | variables |
|  |  | climate |  | are Dunn’s | river basin | for |
|  |  | variables. |  | index, e | and with | prediction |
|  |  | This model |  | Davies– | the above | of rainfall |
|  |  | relates the |  | Bouldin | dataset | states in |
|  |  | large scale |  | index , and | mentioned | future. |
|  |  | variables |  | the | . |  |
|  |  | which are |  | Silhouette |  |  |
|  |  | called as |  | index |  |  |
|  |  | predictors to |  | which |  |  |
|  |  | the local |  | ensures |  |  |
|  |  | scale climate |  | compact- |  |  |
|  |  | which is |  | -ness , |  |  |
|  |  | called as |  | connected |  |  |
|  |  | predictands. |  | ness and |  |  |
|  |  | Transfer |  | separation |  |  |
|  |  | function |  | of cluster |  |  |
|  |  | based method |  | partitions. |  |  |
|  |  | is used in |  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | order to get |  |  |  |  |
|  |  | the linear and |  |  |  |  |
|  |  | non-linear |  |  |  |  |
|  |  | regression |  |  |  |  |
|  |  | methods |  |  |  |  |
|  |  | between the |  |  |  |  |
|  |  | predictors |  |  |  |  |
|  |  | and |  |  |  |  |
|  |  | predictands. |  |  |  |  |
| 22 | Short- | 1)Bayesian |  | In this the | In this | In order to |
|  | series | enhanced |  | computati | particular | improve |
|  | Predictio | modified |  | onal result | paper | the |
|  | n with | approach |  | are is | prediction | performan |
|  | BEMA | (BEMA) |  | evaluated | for | ce of |
|  | Approac | 2) Uses |  | on high | complete | BEMA |
|  | h: | Artificial |  | roughness | and | filter that |
|  | applicati | Neural |  | time series | incomplet | comes |
|  | on to | Network |  | and | e dataset is | under |
|  | short | 3)Non linear |  | compared | establishe | Artifical |
|  | rainfall | filters are |  | with | d and | neural |
|  | series | used |  | artificial | achieved | network |
|  | 2016 |  |  | neural | using this | for |
|  |  |  |  | netwroks | BEMA | comparisi |
|  |  |  |  | and with | combined | on we use |
|  |  |  |  | its non | with | Bayesian |
|  |  |  |  | linear | permutatio | Approach |
|  |  |  |  | filters. | n entropy. | (BA) |
|  |  |  |  |  | The main |  |
|  |  |  |  |  | key area |  |
|  |  |  |  |  | of this |  |
|  |  |  |  |  | paper is |  |
|  |  |  |  |  | incomplet |  |
|  |  |  |  |  | e 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 |  |  |
|  |  |  |  |  |  |  | informatio |  |  |
|  |  |  |  |  |  |  | n. The |  |  |
|  |  |  |  |  |  |  | missing |  |  |
|  |  |  |  |  |  |  | data in the |  |  |
|  |  |  |  |  |  |  | dataset is |  |  |
|  |  |  |  |  |  |  | imputed |  |  |
|  |  |  |  |  |  |  | using |  |  |
|  |  |  |  |  |  |  | linear |  |  |
|  |  |  |  |  |  |  | average |  |  |
|  |  |  |  |  |  |  | smootheni |  |  |
|  |  |  |  |  |  |  | ng. |  |  |
| 23 |  | Modellin |  | 1)Bayesian | Weather | The | Data set is | The |  |
|  |  | g |  | approach | historical data | training | collected | performan |  |
|  |  | Rainfall |  | 2) Data | collected from | dataset is | which | ce of the |  |
|  |  | Predictio |  | intensive | Indian | compared | contains | model can |  |
|  |  | n Using |  | model. | Metrological | with the | 36 | also be |  |
|  |  | Data |  |  | Department(IM | available | attrbutes | improved |  |
|  |  | Mining |  |  | D) Pune. | dataset in | ut of | by |  |
|  |  | Method: |  |  |  | order to | which | designing |  |
|  |  | A |  |  |  | get the | only 7 | the model |  |
|  |  | Bayesian |  |  |  | accuracy. | attributes | for |  |
|  |  | Approac |  |  |  |  | are related | scalable |  |
|  |  | h |  |  |  |  | rainfall | platforms, |  |
|  |  |  |  |  |  |  | prediction. | either for |  |
|  |  | 2013 |  |  |  |  | First the | vertical |  |
|  |  |  |  |  |  |  | data | scalability |  |
|  |  |  |  |  |  |  | undergoes | or for |  |
|  |  |  |  |  |  |  | pre - | horizontal |  |
|  |  |  |  |  |  |  | processing | scalability. |  |
|  |  |  |  |  |  |  | and later |  |  |
|  |  |  |  |  |  |  | data |  |  |
|  |  |  |  |  |  |  | transforma |  |  |
|  |  |  |  |  |  |  | tion is |  |  |
|  |  |  |  |  |  |  | done in |  |  |
|  |  |  |  |  |  |  | order to |  |  |
|  |  |  |  |  |  |  | work on |  |  |
|  |  |  |  |  |  |  | Bayesian. |  |  |
| 24 |  | Monthly |  | 1) Wavelet | Darjeeling | The | Wavelet | This study |  |
|  |  | Rainfall |  | technique | monthly rainfall, | performan | analysis | used only |  |
|  |  | Predictio |  | 2) Artificial | minimum and | ce of | plays a | dataset |  |
|  |  | n Using |  | Neural | maximum | various | major role | from one |  |
|  |  | Wavelet |  | Network | temperature is | models | when | particular |  |
|  |  | Neural |  |  | taken. | during | compared | gague |  |
|  |  | Network |  |  |  | calibration | to all other |  |  |
|  |  | Analysis |  |  |  | and | methods |  |  |
|  |  | 2013 |  |  |  | validation | like |  |  |
|  |  |  |  |  | is are done | Fourier |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  | using | series. |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | various | | | | | | | | | | | |  | Wavelet |  |
|  |  |  |  |  | statistical | | | | | | | | | | | |  | concept |  |
|  |  |  |  |  | indices | | | | | | | | | | | |  | can be |  |
|  |  |  |  |  | they are | | | | | | | | | | | |  | applied to |  |
|  |  |  |  |  | Root | | |  |  |  |  |  |  |  |  |  |  | any size of |  |
|  |  |  |  |  | Mean | | | | |  |  |  |  |  |  |  |  | time |  |
|  |  |  |  |  | Squared | | | | | | |  |  |  |  |  |  | series. |  |
|  |  |  |  |  | Error | | | |  | |  | |  |  |  |  |  | This is |  |
|  |  |  |  |  | (RMSE), | | | | | | | | |  |  |  |  | used to |  |
|  |  |  |  |  | Correlatio | | | | | | | | | |  |  |  | explore, |  |
|  |  |  |  |  | n |  |  | | | |  | |  | | |  |  | de noise, |  |
|  |  |  |  |  | Coefficien | | | | | | | | | | | |  | smoothen |  |
|  |  |  |  |  | t (R) and | | | | | | | |  | | |  |  | time series |  |
|  |  |  |  |  | Coefficien | | | | | | | | | | | |  | which can |  |
|  |  |  |  |  | t of | |  | | | |  | | | | |  |  | be used |  |
|  |  |  |  |  | Efficiency | | | | | | | | | | |  |  | for other |  |
|  |  |  |  |  | (COE). | | | | | |  | | | | | |  | empirical |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | purposes. |  |
| 25 | Rainfall | 1)Regression | Data set is |  | Accuracy | | | | | | | | | | | |  | In this | Other |
|  | Predictio | 2) Artificial | downloaded |  | is 87% | | | | | | | | | | | |  | particular | methods |
|  | n using | Neural | from National |  | Precision | | | | | | | | | | | |  | paper the | can be |
|  | Data | Network | Oceanic and |  | is 98% | | | | | | | | | | | |  | data set is | used and |
|  | Mining | 3) Data | Atmospheric |  | Recall is | | | | | | | | | | | |  | collected | implement |
|  | Techniqu | Mining | Administration |  | 75% | | |  |  |  |  |  |  |  |  |  |  | and pre- | ed as this |
|  | es | techniques. | (NOAA) |  |  |  |  |  |  |  |  |  |  |  |  |  |  | processed | paper has |
|  | 2013 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | using | been used |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | prestd | only two |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | function. | methods |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | While | they are |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | separating | classificati |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | the dataset | on and |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | most of | clustering. |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | the data |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | goes for as |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | training |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | set and |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | very |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | smaller |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | portion |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | will be |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | considered |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | as test dat |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 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 | Research | 1)Seasonal | Dataset is | Performan | Dataset | In this |
|  | on Real- | trend | collected from | ce is | near | paper they |
|  | Time | decompositio | meteorological | measured | Wuhan | have not |
|  | Local | n using Loess | stations. | by | university | proposed |
|  | Rainfall | (STL) | Real time rain | comparing | is taken | the best |
|  | Predictio | 2)Support | fall prediction | SVM and | and | model for |
|  | n Based | vector | near wuhan | Random | computed | the |
|  | on | machine(SV | university. | forest | using star | dataset. |
|  | MEMS | M) |  | where | topology | Dataset |
|  | Sensors | 3)Random |  | SVM | network. | can be run |
|  |  | Forest |  | shows | With the | in Rattle |
|  |  | 4)Back |  | high | help of | and find |
|  |  | Propagation |  | performan | this they | the error |
|  |  | neural |  | ce than | can predict | rates for |
|  |  | Network |  | Random | if rainfalls | all the |
|  |  |  |  | Forest. | in | models. |
|  |  |  |  |  | particular | Whichever |
|  |  |  |  |  | area ,with | model has |
|  |  |  |  |  | that | lower |
|  |  |  |  |  | knowledge | error rate |
|  |  |  |  |  | students | can be |
|  |  |  |  |  | can be | considered |
|  |  |  |  |  | given | as best |
|  |  |  |  |  | informatio | model. |
|  |  |  |  |  | n to carry |  |
|  |  |  |  |  | the |  |
|  |  |  |  |  | umbrella. |  |
| 27 | Rainfall | A deep | Data set is | Mean | The data is | Improve |
|  | predictio | architecture | collected from | Square | extracted | the |
|  | n: A | combining | environmental | Error | and | proposed |
|  | Deep | the use of an | datawarehouse | (MSE) and | divided | architectur |
|  | Learning | autoencoder | administered by | the Root | into | e for light |
|  | approach | and a | IDEA. The | Mean | training | rain |
|  | 2016 | multilayer | dataset contains | Square | dataset | scenarios, |
|  |  | perceptron is | 47 | Error | and testing | use of |
|  |  | used | characteristics | (RMSE) | dataset. | other deep |
|  |  |  | like temperature, |  | Later it | learning |
|  |  |  | humidity, sun |  | undergoes | architectur |
|  |  |  | brightness, |  | into | es like |
|  |  |  | pressure, wind |  | normalisat | deep |
|  |  |  | speed. |  | ion. Vi = | belief |
|  |  |  |  |  | ai − minai | networks |
|  |  |  |  |  | maxai – | based on |
|  |  |  |  |  | minai. | Restricted |
|  |  |  |  |  |  | Boltzmann |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | where ai | Machines |
|  |  |  |  |  | represents | or |
|  |  |  |  |  | a value to | multiple |
|  |  |  |  |  | normalize | stacked |
|  |  |  |  |  | for the i-th | denoising |
|  |  |  |  |  | variable, | autoencod |
|  |  |  |  |  | min ai is | ers |
|  |  |  |  |  | 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 | Modellin | 1)Neural | Rador data is | Performan | Data is | In future |
|  | g and | network | recorded at | ce is | collected | work data |
|  | Predictio | 2)Classificati | KDVN radar | measured | and pre | from other |
|  | n of | on and | station in | using the | processed. | religions |
|  | Rainfall | regression | Davenport, Iowa | following | In this | will be |
|  | Using | tree |  | methods | process | collected |
|  | Radar | 3)Random |  | like | rador | for further |
|  | Reflectiv | forest |  | 1)Mean | images | improvem |
|  | ity Data: | 4)SVM |  | absolute | have been | ent and |
|  | A Data- | 5)Boosted |  | error | collected | validation. |
|  | Mining | tree |  | 2) Mean | out of |  |
|  | Approac | algorithm |  | square | which it |  |
|  | h |  |  | error and | has a |  |
|  | 2013 |  |  | standard | value -99 |  |
|  |  |  |  | deviation. | 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 | Feature | 1)SVM | The weather | Is | In this | Future |
|  | Selection | 2)K nearest | data, which are | measured | they have | work |
|  | for Very | neighbours | collected from | using | generated | would be |
|  | Short- | algorithms | 408 automatic | statistical | new | pre |
|  | Term | 3)Variant K | weather stations | criteria | population | processing |
|  | Heavy | nearest | during the recent | such as | vector | the data |
|  | Rainfall | neighbours | four years from | standard | from the | with |
|  | Predictio | 4) AWS | 2007 to 2010 | error and | original | various |
|  | n |  |  | fitness | population | methods |
|  | Using |  |  | strength. | a mutant | like |
|  | Evolutio |  |  |  | vector is | representat |
|  | nary |  |  |  | obtained | ion |
|  | Computa |  |  |  | by simply | learning, |
|  | tion |  |  |  | selecting | cyclic |
|  | 2014 |  |  |  | two | loess, |
|  |  |  |  |  | random | contrast, |
|  |  |  |  |  | vectors. | and |
|  |  |  |  |  | No the | quantile |
|  |  |  |  |  | mutant | normalizat |
|  |  |  |  |  | vector is | ion |
|  |  |  |  |  | crossed | algorithms |
|  |  |  |  |  | with the |  |
|  |  |  |  |  | original |  |
|  |  |  |  |  | vector and |  |
|  |  |  |  |  | rest are |  |
|  |  |  |  |  | called as |  |
|  |  |  |  |  | trial |  |
|  |  |  |  |  | vectors. |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 30 | Forecasti | 1) Adaptive | Monthly | Root mean | Since the | Data |
|  | ng | neuro fuzzy | climatic data are | square | forecastin | should be |
|  | Rainfall | inference | collected of | error | g of data is | pre |
|  | Using | system. | weather station | (RMSE), | nonlinear | processed |
|  | Adaptive | 2) Back | Ghandhingar | Correlatio | systems | using |
|  | Neuro- | propagation |  | n | ANFIS | different |
|  | Fuzzy | method. |  | Coefficien | model is | methods. |
|  | Inference | 3)Hybrid |  | t (r), | used in |  |
|  | System | method. |  | Coefficien | order to |  |
|  | (ANFIS) | 4)Genetic |  | t of | predict the |  |
|  | 2014 | algorithm |  | Determina | previous |  |
|  |  | 5)FIS |  | tion (R2) | year data. |  |
|  |  |  |  | and | In this we |  |
|  |  |  |  | Discrepan | have mean |  |
|  |  |  |  | cy ratio | temperatur |  |
|  |  |  |  | (D) | e,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. |  |

**Existing Systems:**

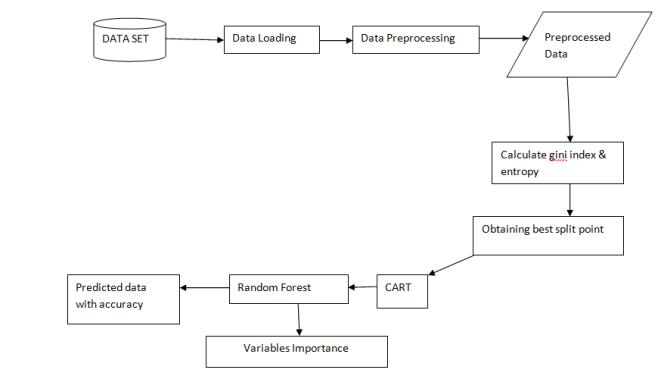
1. In the existing systems, sometimes it happens that the data is not accurate and the proper data mining techniques not being used and this increases the complexity and sometimes the data is of poor quality. But by using proper algorithms for datasets and using the right metrics, we can achieve the accurate results in perdition of rainfall.
2. In the existing system, they have just tried all methods like SVM, Random forest, Neural net but they have not given the best model among them and overall error rate is not calculated.
3. Since the error rate is not calculated, they haven’t tried to reduce the error.

**Gaps Identified:**

1. A decision tree is simple to understand and after a brief exploration, we construct it.
2. It requires modest data training. Earlier techniques demand data normalization, creation of dummy variables and removal of blank values.
3. It can deal with numerical as well as categorical data. Other techniques have their expertise into the analysis of different sets of data having just a single category of variables. Such as neural networks, that can handle only numerical values, while association rules only used with nominal variables.
4. Utilizes the white box model. When a given condition is able to be seen in the model then rationalization for the provided scenario is easily elaborated by Boolean logic.
5. Validation of representation with statistical tests is also possible. It makes the explanation for the reliability of the model possible.
6. It is Robust and generates promising outcomes even when its postulations are debased to some extent with the actual model which generates the data.
7. Time efficient even with large data. Standard computing resources help in analyzing large data.

**Proposed Method:**

**Framework:**



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. The elegant decision tree algorithm has an efficient, accurate and scalable improvement over other algorithms.

Classification and regression techniques are the important methods for decision tree formation. The main advantage of this algorithm is that it decreases the complexity in computation. This helps us in reducing the time taken and random forest is used to increase the accuracy rate. The reason behind selecting this decision tree model is because it has less error rate when compared to other models.

|  |  |  |
| --- | --- | --- |
| S.no | Method name | Overall error rate |
|  |  |  |
| 1 | Decision tree model | 15.6% |
|  |  |  |
| 2 | Linear Model | 17.6% |
|  |  |  |
| 3 | Neural net | 17.6% |
|  |  |  |

**Decision model** is considered as best model since it has less error rate when compared to

others.

**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.

**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:**

Takes random samples and constructs decision trees. Determines the importance of each attribute in the dataset. According to the weight, chooses the best tree. According to the project perspective, the attribute that has higher importance is taken and the bar graph is

plotted. According to the plotted graph, the value of highest frequency is taken as reference for accurate prediction. The dataset is divided into training and testing samples where we apply package (random forest) and test this system using the testing sample.

**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.