Analytical Comparison between the Information Gain and Gini Index using Historical Geographical Data

Dr. Majid Zaman[1]
Directorate of IT & SS
University of Kashmir
Srinagar, India

Sameer Kaul[2]
Department of Computer Science,
University of Kashmir
Srinagar, India

Dr. Muheet Ahmed[3]
Department of Computer Science,
University of Kashmir
Srinagar, India

***Abstract*—The historical geographical data of Kashmir province is spread across two disparate files having attributes of Maximum Temperature, Minimum Temperature, Humidity measured at 12 A.M., Humidity measured at 3 P.M., rainfall besides auxiliary parameters like date, year etc. The parameters Maximum Temperature, Minimum Temperature, Humidity measured at 12 A.M., Humidity measured at 3 P.M. are continuous in nature and here, in this study, we applied Information Gain and Gini Index on these attributes to convert continuous data into discrete values, their after we compare and evaluate the generated results. Of the four attributes, two have same results for Information Gain and Gini Index; one attribute has overlapping results while as only one attribute has conflicting results for Information Gain and Gini Index. Subsequently, continuous valued attributes are converted into discrete values using Gini index. Irrelevant attributes are not considered and auxiliary attributes are labeled accordingly. Consequently, the data set is ready for the application of machine learning (decision tree) algorithms.**

Keywords*—Geographical data mining; information gain; Gini index; machine learning; decision tree

I. Introduction

A. Splitting Rules

Decision tree is built by recursively splitting data partitions into smaller partitions according to splitting rules or criteria. Attribute selection measure or splitting rules is a heuristic for choice of criteria that best splits class labeled training dataset into separate classes. Attribute selection measure should be such that split should produce pure partitions i.e. all the records in given partition belong to same class.

The attribute selection measure gives a score/value for each attribute, best describing given class labeled training dataset, the attribute having best score/value is chosen as splitting attribute for given partition. In this paper we have used Information Gain for the attribute selection measure.

B. Information Gain and Gini Index

ID3 uses information gain as its attribute selection measure. For a given node that holds tuples of partition D, the attribute with highest information gain (score/value) is chosen as splitting attribute for the given node [1][6]. The chosen attribute requires least information for classifying records in the resultant partitions besides discloses least impurity in these partitions, thus resulting in minimum number of tests required to classify a given record and generation of (simple) decision tree, accordingly information required for classification of a record in D is given by (1).

$$\mathrm{Info}(D) = -\sum_{i=1}^{m} pi \log 2(pi) \text{ [5]} \quad (1)$$

and Information still required to arrive at an exact classification is measured by (2).

$$\mathrm{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * \mathrm{Info}(D) \text{ [5]} \quad (2)$$

Information Gain is the difference between the original information requirement and the new requirement, that is

$$\mathrm{Gain}(A) = \mathrm{Info}(D) - \mathrm{InfoA}(D) \text{ [5]} \quad (3)$$

Thus, Gain(A) is the gain if A is chosen for branching, accordingly Gain is calculated for all the attributes of the training set and attribute with the highest information gain is chosen as splitting attribute for the given node[2][3][7]. Thus calculation of information gain enables us to choose the attribute that would do the best classification, further most the amount of information still required for classifying records is minimal.

The Gini Index is used by CART. The Gini index measures the impurity in D[10][11]. The Gini index considers binary split for each attribute; accordingly weighted sum of impurity of each resulting partition is calculated, thus binary split on A partitions D into D1 & D2 i.e. [5].

$$\mathrm{Gini}(D) = 1 - \sum_{i=1}^{m} pi^2 \quad (4)$$

$$\text{and } \mathrm{GiniA}(D) = \frac{|D_1|}{|D|}\mathrm{Gini}(D1) + \frac{|D_2|}{|D|}\mathrm{Gini}(D2) \quad (5)$$

The reduction in impurity that would be incurred by a binary split on a discrete on attribute A is

$$\mathrm{Gini}(A) = \mathrm{Gini}(D) - \mathrm{GiniA}(D) \quad (6)$$

The process is repeated for every attribute and the attribute that has minimum Gini index is chosen as splitting attribute [2][3][8].

C. Continuous Valued Attributes

For an attribute "A" that has continuous values e.g. temperature, humidity etc. the best split point is to be determined for "A". All the possible unique values of A are sorted in ascending order, the midpoint between two adjacent values is considered [5].

$$\frac{ai+ai+1}{2} \quad (7)$$

for the given unique u values of attribute A, u-1 values will be generated, for each generated value infoA(D) is calculated with number of partitions two [4][9][12] .The mid-point with minimum value is chosen as the split point of A where

D1 is set of records satisfying

D2 is set of records satisfying

The other possible solution is to calculate Gini index for every mid-point (Gini index is calculated instead of infoA (D)) and minimum Gini index for a give attribute is taken as split point of the attribute.

II. Related Work

Gini index and Information gain have been used extensively used over the years, however most relevant work done in the recent past on the comparison of Gini index and Information gain is presented below.

In their research paper entitled "Theoretical comparison between the Gini Index and Information Gain criteria" Laura Elena Raileanu and Kilian Stoffel proposed a formal methodology to compare multiple split criteria and also presented a formal description of how to select between split criteria for a given data set, they concluded that Information Gain and Gini Index disagree only in 2% of all cases [13].

Mohammed A. Muharram and George D. Smith compared the performance of classifiers in their paper "Evolutionary Feature Construction Using Information Gain and Gini Index" to ascertain if C5 or CART was in any way benefiting from the inclusion of an attribute evolved using Information gain or Gini index respectively, they found no evidence that any algorithm has an advantage over the other classifiers and according to them all classifiers benefit from the inclusion of an evolved attribute [14].

Theoretical and empirical comparison of different split measures for induction of decision tree in Random forest and its effect on the accuracy of Random forest was done by Vrushali Y. Kulkarni, Manisha Petare and P. K. Sinha in their work entitled Analyzing Random Forest Classifier with Different Split Measures. The empirical results put forth by them, show that there is not much / significant variation in accuracy obtained except Chi Square, further Information gain and Gain ratio give comparable results for almost all datasets and Gini index slightly lags in the results with most of the datasets [15].

III. Data

The data used in this paper is split across two CSV files, which has been collected from NDC Pune (India Meteorological department), agency of Ministry of earth sciences, Government of India. It is the principal agency responsible for meteorological observations, weather forecasting and seismology. IMD is one of the six regional specialized meteorological centers of the world meteorological organization.

The weather parameters in both data files are taken for the 3 regions of Kashmir division i.e. Gulmarg (North Kashmir), Srinagar (Central Kashmir) and Qazigund (South Kashmir). Gulmarg is geographically located at 34.05°N 74.38°E and has an average elevation of 2,650 m (8,690 ft.), Srinagar (Central) is located at 34.5°N 74.47°E and has an average elevation of 1,585 m (5,200 ft.), and Qazigund (South) is located at 33.59°N 75.16°E. It has an average elevation of 1,670 m (5,480 ft.).

The first data file (Fig. 1), shown below consists of 12190 instances of relative humidity (in %) measured every day at time 12 AM and 3 PM from year 2012 to 2017, for all the three stations.

The second data file (Fig. 2), shown below consists of 6117 instances of Maximum temperature (°C), Minimum temperature (°C) and Rainfall (in mm) measured every day from year 2012 to 2017, for all the three stations.

The two data files are integrated into single holistic dataset, discrepancies are resolved, data for each attribute is cleaned, transformed and loaded for formation of single dataset, shown below (Fig 3). The integrated data has Maximum temperature (tmax), Minimum temperature (tmin) and Rainfall (rfall), humidity measured 12 AM (humid12) and 3 PM (humid3) for every day (with exception) from year 2012 to 2017, for all the three stations.

station_id	year	mnth	hr	dt	rhumid
42026	2012	1	3	1	100
42026	2012	1	3	2	100
42026	2012	1	3	3	96
42026	2012	1	3	4	100
42026	2012	1	3	5	100
42026	2012	1	3	6	100
42026	2012	1	3	7	100
42026	2012	1	3	8	100
42026	2012	1	3	9	100
42026	2012	1	3	10	86
42026	2012	1	3	11	87
42026	2012	1	3	12	100
42026	2012	1	3	13	100
42026	2012	1	3	14	100
42026	2012	1	3	15	100
42026	2012	1	3	16	100
42026	2012	1	3	17	100

Fig. 1. Instances of Relative Humidity at 12 am and 3pm.

station_id	year	mnth	dt	tmax	tmin	rfall
42026	2012	1	1	5.5	-8	0
42026	2012	1	2	5.4	-7.6	0
42026	2012	1	3	4.2	-8	0
42026	2012	1	4	4	-7.2	0
42026	2012	1	5	-1	-9.1	1.1
42026	2012	1	6	-2	-8	17.9
42026	2012	1	7	-1	-10.5	6.8
42026	2012	1	8	1	-16.5	12.6
42026	2012	1	9	-2.8	-14.5	0
42026	2012	1	10	-2.5	-16.2	0
42026	2012	1	11	-7.8	-14.8	0
42026	2012	1	12	-8.2	-16.4	0
42026	2012	1	13	-7.5	-16.5	0
42026	2012	1	14	-7.5	-15.2	0
42026	2012	1	15	-1.5	-9.6	16
42026	2012	1	16	-3	-6.7	21

Fig. 2. Instances of Maximum Temperature, Minimum Temperature and Rainfall.

station_id	year	mnth	dt	tmax	tmin	rfall	humid3	humid12
42026	2012	1	1	5.5	-8	0	100	100
42026	2012	1	2	5.4	-7.6	0	100	100
42026	2012	1	3	4.2	-8	0	96	90
42026	2012	1	4	4	-7.2	0	100	100
42026	2012	1	5	-1	-9.1	1.1	100	100
42026	2012	1	6	-2	-8	17.9	100	100
42026	2012	1	7	-1	-10.5	6.8	100	100
42026	2012	1	8	1	-16.5	12.6	100	100
42026	2012	1	9	-2.8	-14.5	0	100	83
42026	2012	1	10	-2.5	-16.2	0	86	94
42026	2012	1	11	-7.8	-14.8	0	87	100
42026	2012	1	12	-8.2	-16.4	0	100	100
42026	2012	1	13	-7.5	-16.5	0	100	100
42026	2012	1	14	-7.5	-15.2	0	100	100
42026	2012	1	15	-1.5	-9.6	16	100	100

Fig. 3. Cleaned and Integrated Dataset.

A. Data Attributes

Of the nine attributes five are geographical parameters, they are Maximum Temperature, Minimum Temperature, Rainfall, Humidity at 12 & Humidity at 3 termed as tmax, tmin, rfall, humid12 & humid3 respectively, while as four parameters are auxiliary/dependent parameters they are station id, year, month and date termed as station_id, year, mnth & dt. In order to implement decision tree for the prediction of rainfall we have to evaluate each attribute of the resultant data independently.

1) Rainfall: As per the resultant dataset the rainfall in Kashmir province varies from no rainfall to above 100 mm of rainfall in one day. The broader inspection of rain data of five years recorded in 5951 entries is that there is no rainfall in 4026 instances and rainfall in 1952 instances, thus the inference is that we can divide rain data in to two classes that is presence and absence of rain, accordingly dataset is to be modified with new column "crfall" which will be marked as "Y" in case of rainfall (1925 entries) and "N" in case of no rainfall (4026 entries). The Decision Tree is trained to predict presence or absence of rain on a given day.

2) Maximum temperature: Maximum Temperature (tmax) is continuous valued rather than discrete valued, in this case we must determine the "best" split-point for Maximum Temperature (tmax), where the split-point is a threshold on Maximum Temperature (tmax), this can be determined by employing either of the two techniques, Information Gain used by ID3 or Gini Index used by CART, in this paper we use both the techniques to determine the split-point, we will compare the results from the two techniques (Information Gain & Gini Index) and decide accordingly. In order to calculate Information Gain or Gini index, we need to determine unique values of Maximum Temperature (tmax) and then these unique values are to be sorted in ascending order. In the dataset of 5951 records there are 380 unique values of Maximum Temperature (tmax) recoded, varying from -8.2°C to 35.4°C. Their after mid-point between each pair of adjacent values is considered as possible split-point., the snap shot of first 10, middle 10 and last 10 sorted records with mid points are shown in Fig. 4.

rno	tmax	spltpnt
1	-8.2	0
2	-7.8	-8
3	-7.6	-7.7
4	-7.5	-7.55
5	-6.7	-7.1
6	-5.5	-6.1
7	-5	-5.25
8	-4.5	-4.75
9	-4.4	-4.45
10	-4	-4.2
141	10.6	10.55
142	10.7	10.65
143	10.8	10.75
144	10.9	10.85
145	11	10.95
146	11.1	11.05
147	11.2	11.15
148	11.3	11.25
149	11.4	11.35
150	11.5	11.45
371	33.6	33.55
372	33.7	33.65
373	33.8	33.75
374	33.9	33.85
375	34	33.95
376	34.1	34.05
377	34.2	34.15
378	34.4	34.3
379	34.6	34.5
380	35.4	35

Fig. 4. Unique Values of Maximum Temperatures and their Split Points.

Therefore given 380 values of Maximum Temperature (tmax), 379 possible splits will be evaluated, accordingly there shall be no mid-point generated for first recoded temperature -8.2°C because there is no prior temperature value. For example, the mid-point between the values of 33.8 and 33.9 of Maximum Temperature (rno 373 & 374) is 33.85, which is listed in the table for rno 374 against the value of 33.9.

$$\frac{33.8 + 33.9}{2} = 33.85$$

For each possible split-point for Maximum Temperature, we will evaluate $Info_{tmax}(D)$ and $Gini_{tmax}(D)$ but first we have to determine the prerequisites, for possible split value of 33.85 we have to determine the following:

1) fyes: No. of days there was rain for tmax<=33.85
2) fno: No. of days there was no rain for tmax<=33.85
3) syes: No. of days there was rain for tmax>33.85
4) sno: No. of days there was no rain for tmax>33.85

These values have to be generated for all possible split-points, the snap shot of first 10, middle 10 and the last 10 records with necessary values are shown below (Fig. 5).

Again first row shall not be considered because it has no mid-point, for every other possible point we have generated necessary values.

For each possible split-point for Maximum Temperature, we will calculate $Info_{spltpnt}(D)$ and $Gini_{spltpnt}(D)$ using following equations

$$Info(D) = -\sum_{i=1}^{m} pi\ log2(pi)\ [5] \quad (8)$$

$$Info_{spltpnt}(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * Info(D) \quad (9)$$

and $$Gini(D) = 1 - \sum_{i=1}^{m} pi^2 \quad (10)$$

$$Gini_{spltpnt}(D) = \frac{|D_1|}{|D|} Gini(D1) + \frac{|D_2|}{|D|} Gini(D2) \quad (11)$$

rno	tmax	spltpnt	fyes	fno	syes	sno
1	-8.2	0	0	0	0	0
2	-7.8	-8	0	1	1925	4025
3	-7.6	-7.7	0	2	1925	4024
4	-7.5	-7.55	0	3	1925	4023
5	-6.7	-7.1	0	5	1925	4021
6	-5.5	-6.1	0	6	1925	4020
7	-5	-5.25	1	7	1924	4019
8	-4.5	-4.75	2	7	1923	4019
9	-4.4	-4.45	3	7	1922	4019
10	-4	-4.2	3	8	1922	4018
141	10.6	10.55	647	858	1278	3168
142	10.7	10.65	653	872	1272	3154
143	10.8	10.75	655	876	1270	3150
144	10.9	10.85	665	887	1260	3139
145	11	10.95	666	896	1259	3130
146	11.1	11.05	674	922	1251	3104
147	11.2	11.15	674	933	1251	3093
148	11.3	11.25	678	951	1247	3075
149	11.4	11.35	678	965	1247	3061
150	11.5	11.45	681	983	1244	3043
371	33.6	33.55	1922	4005	3	21
372	33.7	33.65	1923	4008	2	18
373	33.8	33.75	1923	4009	2	17
374	33.9	33.85	1923	4013	2	13
375	34	33.95	1923	4017	2	9
376	34.1	34.05	1924	4019	1	7
377	34.2	34.15	1925	4021	0	5
378	34.4	34.3	1925	4022	0	4
379	34.6	34.5	1925	4024	0	2
380	35.4	35	1925	4025	0	1

Fig. 5. Possible Splitpoints for Maximum Temperature.

For example, we will generate Info(D) for a possible split-point of 10.85 listed in above table for rno 144 for tmax of 10.9.

$$Info_{spltpnt}(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * Info(D) \quad (12)$$

$$Info_{10.9}(D) = \frac{1552}{5951} * \left(-\frac{665}{1552} * \text{LOG2}\left(\frac{665}{1552}\right) - \frac{887}{1552} * \text{LOG2}\left(\frac{887}{1552}\right)\right) + \frac{4399}{5951} * \left(-\frac{1260}{4399} * \text{LOG2}\left(\frac{1260}{4399}\right) - \frac{3139}{4399} * \text{LOG2}\left(\frac{3139}{4399}\right)\right)$$

$Info_{10.9}(D)$= 0.895652633

And we will generate Gini(D) for a possible split-point of 10.85 listed in above table for rno 144 for tmax of 10.9.

$$\text{Gini}_{\text{spltpnt}}(D) = \frac{|D_1|}{|D|} Gini(D1) + \frac{|D_2|}{|D|} Gini(D2) \quad [5]$$

$$\text{Gini}_{10.9}(D) = \frac{1552}{5951} * \left(1 - \left(\frac{665}{1552}\right)^2 - \left(\frac{887}{1552}\right)^2\right) + \frac{4399}{5951} * \left(1 - \left(\frac{1260}{4399}\right)^2 - \left(\frac{3139}{4399}\right)^2\right)$$

$\text{Gini}_{10.9}(D)$= 0.429897835

Likewise we generate Info(D) and Gini(D) for each possible split-point for Maximum Temperature, the snap shot of first 10, middle 10 and last 10 records with necessary values are shown in Fig. 6.

In this way we generate Info(D) and Gini(D) for every possible split-point, with exception to rno 1 because it has no split point, further of 379 possible split-points 9 possible split-points do not generate info(D), show below (Fig. 7).

This is because one of the values of fyes, fno, syes, sno is zero. We have generated Information Gain and Gini Index for every split point; we now compare the two results.

Case 1: Information Gain

The point with minimum expected information requirement for Maximum Temperature (tmax) is to be selected as the split point for Maximum Temperature (tmax), the five best cases with minimum Information Gain are shown below (Fig. 8).

The above table is regenerated with Gini Index for the above split-points (Fig 9).

rno	tmax	spltpnt	fyes	fno	syes	sno	info	gini
1	-8.2	0	0	0	0	0	0	0
2	-7.8	-8	0	1	1925	4025	0	0.437643
3	-7.6	-7.7	0	2	1925	4024	0	0.437608
4	-7.5	-7.55	0	3	1925	4023	0	0.437572
5	-6.7	-7.1	0	5	1925	4021	0	0.437502
6	-5.5	-6.1	0	6	1925	4020	0	0.437467
7	-5	-5.25	1	7	1924	4019	0.907914	0.437572
8	-4.5	-4.75	2	7	1923	4019	0.908066	0.437647
9	-4.4	-4.45	3	7	1922	4019	0.908117	0.437676
10	-4	-4.2	3	8	1922	4018	0.908104	0.437668
141	10.6	10.55	647	858	1278	3168	0.895845	0.43001
142	10.7	10.65	653	872	1272	3154	0.896015	0.430121
143	10.8	10.75	655	876	1270	3150	0.896035	0.430134
144	10.9	10.85	665	887	1260	3139	0.895653	0.429898
145	11	10.95	666	896	1259	3130	0.896035	0.430141
146	11.1	11.05	674	922	1251	3104	0.896625	0.430519
147	11.2	11.15	674	933	1251	3093	0.897178	0.430868
148	11.3	11.25	678	951	1247	3075	0.897694	0.431196
149	11.4	11.35	678	965	1247	3061	0.898353	0.431611
150	11.5	11.45	681	983	1244	3043	0.898915	0.431966
371	33.6	33.55	1922	4005	3	21	0.9075	0.437359
372	33.7	33.65	1923	4008	2	18	0.907442	0.437341
373	33.8	33.75	1923	4009	2	17	0.907511	0.437373
374	33.9	33.85	1923	4013	2	13	0.907769	0.437495
375	34	33.95	1923	4017	2	9	0.907985	0.437604
376	34.1	34.05	1924	4019	1	7	0.907914	0.437572
377	34.2	34.15	1925	4021	0	5	0	0.437502
378	34.4	34.3	1925	4022	0	4	0	0.437537
379	34.6	34.5	1925	4024	0	2	0	0.437608
380	35.4	35	1925	4025	0	1	0	0.437643

Fig. 6. Information Gain and Gini for each Possible Split-Point for Maximum Temperature.

rno	tmax	spltpnt	fyes	fno	syes	sno	info	gini
2	-7.8	-8	0	1	1925	4025	0	0.437643
3	-7.6	-7.7	0	2	1925	4024	0	0.437608
4	-7.5	-7.55	0	3	1925	4023	0	0.437572
5	-6.7	-7.1	0	5	1925	4021	0	0.437502
6	-5.5	-6.1	0	6	1925	4020	0	0.437467
377	34.2	34.15	1925	4021	0	5	0	0.437502
378	34.4	34.3	1925	4022	0	4	0	0.437537
379	34.6	34.5	1925	4024	0	2	0	0.437608
380	35.4	35	1925	4025	0	1	0	0.437643

Fig. 7. Split-Points where Information Gain is not Generated for Maximum Temperature.

rno	spltpnt	info
286	25.05	0.891764
285	24.95	0.892011
284	24.85	0.892055
288	25.25	0.892174
283	24.75	0.892284

Fig. 8. Five Best Cases with Minimum Information Gain for Maximum Temperature.

rno	spltpnt	info	Gini
286	25.05	0.891764	0.428311
285	24.95	0.892011	0.428436
284	24.85	0.892055	0.428453
288	25.25	0.892174	0.428548
283	24.75	0.892284	0.428572

Fig. 9. Gini Index for each Respected Split-Point.

and in accordance to the rule of Information Gain we have to choose 25.05 as split-point for Maximum Temperature(tmax) since it has the lowest Information Gain, split-point 25.05 with all the attributes is shown below: (Fig. 10).

Case 2: Gini Index

The point giving the minimum Gini index for a given attribute Maximum Temperature (tmax) is to be taken as a split-point for the Maximum Temperature (tmax), the five best cases with minimum Gini Index are shown below: (Fig. 11).

The above table is regenerated with Information Gain for the above split-points (Fig. 12).

And in accordance to the rule we have to choose 8.05 as split-point for Maximum Temperature(tmax) since it has the lowest Gini Index, split-point 8.05 with all the attributes is shown below (Fig. 13).

rno	tmax	spltpnt	fyes	fno	syes	sno	info	gini
286	25.1	25.05	1624	2853	301	1173	0.891764	0.428311

Fig. 10. Split-Point with Lowest Information Gain for Maximum Temperature.

rno	spltpnt	info
116	8.05	0.893182
120	8.45	0.893214
113	7.75	0.893338
112	7.65	0.893355
121	8.55	0.893285

Fig. 11. Five Best Cases with Minimum Gini Index for Maximum Temperature.

rno	spltpnt	gini	info
116	8.05	0.428188	0.893182
120	8.45	0.42823	0.893214
113	7.75	0.428271	0.893338
112	7.65	0.42828	0.893355
121	8.55	0.428284	0.893285

Fig. 12. Information Gain for Each Respected Split-Point.

rno	tmax	spltpnt	fyes	fno	syes	sno	info	gini
116	8.1	8.05	494	551	1431	3475	0.893182	0.428188

Fig. 13. Split-Point with Lowest Gini Index for Maximum Temperature.

The results of Information Gain and Gini Index do not corroborate, and hence we have to choose one of the values, either as per Information Gain (25.05) or as per Gini Index (8.05).

3) Minimum Temperature: Minimum Temperature (tmin) is again continuous valued rather than discrete valued, in this case we must determine the "best" split-point for Minimum Temperature (tmin), where the split-point is a threshold on Minimum Temperature (tmin), again we use both the techniques to determine the split-point, we will compare the results from the two techniques (Information Gain & Gini Index) and decide accordingly.

We determine unique values of Minimum Temperature (tmin) and then these unique values are sorted in ascending order. In the dataset of 5951 records there are 354 unique values of Minimum Temperature (tmin) recoded, varying from -16.5°C to 23.8°C. Their after mid-point between each pair of adjacent values is generated as possible split-point., the snap shot of first 10, middle 10 and last 10 sorted records with mid points are shown below (Fig. 14).

Therefore given 354 values of Minimum Temperature (tmin), 353 possible splits will be generated and evaluated, there is no mid-point generated for the first minimum recorded temperature -16.5°C.

rno	tmin	spltpnt
1	-16.5	0
2	-16.4	-16.45
3	-16.2	-16.3
4	-15.2	-15.7
5	-14.8	-15
6	-14.6	-14.7
7	-14.5	-14.55
8	-14.4	-14.45
9	-14.2	-14.3
10	-14	-14.1
178	5.1	5.05
179	5.2	5.15
180	5.3	5.25
181	5.4	5.35
182	5.5	5.45
183	5.6	5.55
184	5.7	5.65
185	5.8	5.75
186	5.9	5.85
187	6	5.95
345	21.8	21.75
346	21.9	21.85
347	22	21.95
348	22.2	22.1
349	22.5	22.35
350	22.6	22.55
351	22.8	22.7
352	22.9	22.85
353	23.1	23
354	23.8	23.45

Fig. 14. Unique Values of Minimum Temperatures and their Split Points.

For each possible split-point for Minimum Temperature (tmin), we calculate values of fyes, fno, syes, and sno. These values have to be generated for all possible split-points, the snap shot of first 10, middle 10 and last 10 records with necessary values are shown below (Fig. 15).

Again first row shall not be considered because it has no mid-point, for every other possible point we have generated necessary values.

For each possible split-point for Minimum Temperature, we will calculate $Info_{spltpnt}(D)$ and $Gini_{spltpnt}(D)$ using following equations. (13)(14)(15)(16).

$$Info(D) = -\sum_{i=1}^{m} pi\ log2(pi) \quad (13)$$

$$Info_{spltpnt}(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * Info(D) \quad (14)$$

and $$Gini(D) = 1 - \sum_{i=1}^{m} pi^2 \quad (15)$$

$$Gini_{spltpnt}(D) = \frac{|D_1|}{|D|} Gini(D1) + \frac{|D_2|}{|D|} Gini(D2) \quad (16)$$

The snap shot of first 10, middle 10 and last 10 records with necessary values are shown in Fig. 16.

We generate Info(D) and Gini(D) for every possible split-point with exception to rno 1 because it has no split point, further of 353 possible split-points 12 possible split-points do not generate info(D), this is because one of the values of fyes, fno, syes, sno is zero, as shown in Fig. 17.

We have generated Information Gain and Gini Index for every split point; we now compare the two results.

rno	tmin	spltpnt	fyes	fno	syes	sno
1	-16.5	0	0	0	0	0
2	-16.4	-16.45	1	1	1924	4025
3	-16.2	-16.3	1	2	1924	4024
4	-15.2	-15.7	1	4	1924	4022
5	-14.8	-15	1	5	1924	4021
6	-14.6	-14.7	1	6	1924	4020
7	-14.5	-14.55	3	6	1922	4020
8	-14.4	-14.45	3	7	1922	4019
9	-14.2	-14.3	3	8	1922	4018
10	-14	-14.1	3	9	1922	4017
178	5.1	5.05	837	1978	1088	2048
179	5.2	5.15	839	1981	1086	2045
180	5.3	5.25	848	2004	1077	2022
181	5.4	5.35	853	2011	1072	2015
182	5.5	5.45	866	2027	1059	1999
183	5.6	5.55	871	2038	1054	1988
184	5.7	5.65	881	2057	1044	1969
185	5.8	5.75	888	2073	1037	1953
186	5.9	5.85	897	2093	1028	1933
187	6	5.95	900	2098	1025	1928
345	21.8	21.75	1925	4002	0	24
346	21.9	21.85	1925	4007	0	19
347	22	21.95	1925	4013	0	13
348	22.2	22.1	1925	4014	0	12
349	22.5	22.35	1925	4018	0	8
350	22.6	22.55	1925	4020	0	6
351	22.8	22.7	1925	4021	0	5
352	22.9	22.85	1925	4022	0	4
353	23.1	23	1925	4024	0	2
354	23.8	23.45	1925	4025	0	1

Fig. 15. Possible Splitpoints for Minimum Temperature.

rno	tmin	spltpnt	fyes	fno	syes	sno	info	gini
1	-16.5	0	0	0	0	0	0	0
2	-16.4	-16.45	1	1	1924	4025	0.908088	0.437657
3	-16.2	-16.3	1	2	1924	4024	0.90812	0.437678
4	-15.2	-15.7	1	4	1924	4022	0.908074	0.437652
5	-14.8	-15	1	5	1924	4021	0.908028	0.437628
6	-14.6	-14.7	1	6	1924	4020	0.907974	0.437601
7	-14.5	-14.55	3	6	1922	4020	0.90812	0.437678
8	-14.4	-14.45	3	7	1922	4019	0.908117	0.437676
9	-14.2	-14.3	3	8	1922	4018	0.908104	0.437668
10	-14	-14.1	3	9	1922	4017	0.908083	0.437656
178	5.1	5.05	837	1978	1088	2048	0.906094	0.436451
179	5.2	5.15	839	1981	1086	2045	0.906116	0.436464
180	5.3	5.25	848	2004	1077	2022	0.906044	0.43642
181	5.4	5.35	853	2011	1072	2015	0.906107	0.436458
182	5.5	5.45	866	2027	1059	1999	0.906302	0.436576
183	5.6	5.55	871	2038	1054	1988	0.906294	0.436571
184	5.7	5.65	881	2057	1044	1969	0.906327	0.436591
185	5.8	5.75	888	2073	1037	1953	0.906305	0.436577
186	5.9	5.85	897	2093	1028	1933	0.906285	0.436565
187	6	5.95	900	2098	1025	1928	0.906307	0.436578
345	21.8	21.75	1925	4002	0	24	0	0.43683
346	21.9	21.85	1925	4007	0	19	0	0.437008
347	22	21.95	1925	4013	0	13	0	0.43722
348	22.2	22.1	1925	4014	0	12	0	0.437255
349	22.5	22.35	1925	4018	0	8	0	0.437396
350	22.6	22.55	1925	4020	0	6	0	0.437467
351	22.8	22.7	1925	4021	0	5	0	0.437502
352	22.9	22.85	1925	4022	0	4	0	0.437537
353	23.1	23	1925	4024	0	2	0	0.437608
354	23.8	23.45	1925	4025	0	1	0	0.437643

Fig. 16. Information Gain and Gini for each Possible Split-Point for Minimum Temperature.

rno	tmin	spltpnt	fyes	fno	syes	sno	info	gini
343	21.6	21.55	1925	3994	0	32	0	0.436546
344	21.7	21.65	1925	3997	0	29	0	0.436653
345	21.8	21.75	1925	4002	0	24	0	0.43683
346	21.9	21.85	1925	4007	0	19	0	0.437008
347	22	21.95	1925	4013	0	13	0	0.43722
348	22.2	22.1	1925	4014	0	12	0	0.437255
349	22.5	22.35	1925	4018	0	8	0	0.437396
350	22.6	22.55	1925	4020	0	6	0	0.437467
351	22.8	22.7	1925	4021	0	5	0	0.437502
352	22.9	22.85	1925	4022	0	4	0	0.437537
353	23.1	23	1925	4024	0	2	0	0.437608
354	23.8	23.45	1925	4025	0	1	0	0.437643

Fig. 17. Split-Points where Information Gain is not Generated for Minimum Temperature.

Case 1: Information Gain

The point with minimum expected information requirement for Minimum Temperature (tmin) is to be selected as the split point for Minimum Temperature (tmin), the five best cases with minimum Information Gain are shown below: (Fig 18).

The above table is regenerated with Gini Index for the split-points (Fig. 19).

rno	spltpnt	info
124	-0.35	0.900033
125	-0.25	0.900106
123	-0.45	0.900434
122	-0.55	0.900468
120	-0.75	0.900554

Fig. 18. Five Best cases with Minimum Information Gain for Minimum Temperature.

rno	spltpnt	info	gini
124	-0.35	0.900033	0.432954
125	-0.25	0.900106	0.432992
123	-0.45	0.900434	0.43319
122	-0.55	0.900468	0.433212
120	-0.75	0.900554	0.433268

Fig. 19. Gini Index for each Respected Split-Point.

And in accordance to the rule of Information Gain we have to choose -0.35 as split-point for Minimum Temperature (tmin) since it has the lowest Information Gain, split-point -0.35 with all the attributes is shown below: (Fig 20).

Case 2: Gini Index

The point giving the minimum Gini index for a given attribute Minimum Temperature (tmin) is to be taken as a split-point for the Minimum Temperature (tmin), the five best cases with minimum Gini Index are shown below: (Fig. 21).

The table is regenerated with Information Gain for the above split-points: (Fig. 22).

And in accordance to the rule of Gini Index we have to choose -0.35 as split-point for Minimum Temperature (tmin) since it has the lowest Gini Index, split-point -0.35 with all the attributes is shown below: (Fig. 23).

The results of Information Gain and Gini Index are exactly the same, hence split-point -0.35 will be chosen in either case, and there is no conflict at all.

rno	tmin	spltpnt	fyes	fno	syes	sno	info	gini
124	-0.3	-0.35	349	1115	1576	2911	0.900033	0.432954

Fig. 20. Split-Point with Lowest Information Gain for Minimum Temperature.

rno	spltpnt	gini
124	-0.35	0.432954
125	-0.25	0.432992
123	-0.45	0.43319
122	-0.55	0.433212
120	-0.75	0.433268

Fig. 21. Five best cases with Minimum Gini Index for Minimum Temperature.

rno	spltpnt	gini	info
124	-0.35	0.432954	0.900033
125	-0.25	0.432992	0.900106
123	-0.45	0.43319	0.900434
122	-0.55	0.433212	0.900468
120	-0.75	0.433268	0.900554

Fig. 22. Information Gain for each Respected Split-Point.

rno	tmin	spltpnt	fyes	fno	syes	sno	info	gini
124	-0.3	-0.35	349	1115	1576	2911	0.900033	0.432954

Fig. 23. Split-point with lowest Gini Index for Minimum Temperature.

4) Humidity Measured at 12:00 A.M: Like Maximum Temperature (tmax) & Minimum Temperature (tmin) Humidity Measured at 12:00 A.M (humid12) is continuous valued rather than discrete valued, and in accordance with the methodology used for the determination of best split-point for maximum and minimum temperature, we use same procedure for determination of best split-point for humidity12 as well. In the dataset of 5951 records there are 82 unique values of Humidity Measured at 12:00 A.M (humid12) recoded, varying from 18 to 100. The snap shot of first 10, middle 10 and last 10-sorted records with mid points are shown below (Fig. 24), 81 possible split-points will be evaluated.

The snap shot of first 10, middle 10 and last 10 records with necessary values of fyes, fno, syes & sno are shown below (Fig. 25).

rno	humid12	spltpnt
1	18	0
2	19	18.5
3	20	19.5
4	21	20.5
5	22	21.5
6	23	22.5
7	24	23.5
8	25	24.5
9	26	25.5
10	27	26.5
42	59	58.5
43	60	59.5
44	61	60.5
45	62	61.5
46	63	62.5
47	64	63.5
48	65	64.5
49	66	65.5
50	67	66.5
51	68	67.5
73	90	89.5
74	91	90.5
75	92	91.5
76	93	92.5
77	94	93.5
78	95	94.5
79	96	95.5
80	97	96.5
81	98	97.5
82	100	99

Fig. 24. Unique Values of Humidity at 12 am and their Split Points.

rno	humid12	spltpnt	fyes	fno	syes	sno
1	18	0	0	0	0	0
2	19	18.5	0	1	1925	4025
3	20	19.5	0	5	1925	4021
4	21	20.5	0	7	1925	4019
5	22	21.5	0	14	1925	4012
6	23	22.5	0	23	1925	4003
7	24	23.5	0	37	1925	3989
8	25	24.5	0	47	1925	3979
9	26	25.5	1	57	1924	3969
10	27	26.5	1	75	1924	3951
42	59	58.5	452	2341	1473	1685
43	60	59.5	483	2437	1442	1589
44	61	60.5	516	2543	1409	1483
45	62	61.5	551	2609	1374	1417
46	63	62.5	583	2699	1342	1327
47	64	63.5	600	2755	1325	1271
48	65	64.5	630	2817	1295	1209
49	66	65.5	653	2880	1272	1146
50	67	66.5	685	2944	1240	1082
51	68	67.5	711	3003	1214	1023
73	90	89.5	1406	3802	519	224
74	91	90.5	1437	3825	488	201
75	92	91.5	1493	3852	432	174
76	93	92.5	1529	3877	396	149
77	94	93.5	1578	3894	347	132
78	95	94.5	1622	3907	303	119
79	96	95.5	1643	3917	282	109
80	97	96.5	1682	3944	243	82
81	98	97.5	1731	3955	194	71
82	100	99	1741	3957	184	69

Fig. 25. Possible Splitpoints for Humidity at 12 am.

For each possible split-point for Minimum Temperature, we will calculate $Info_{spltpnt}(D)$ and $Gini_{spltpnt}(D)$ using following equations.(17)(18)(19)(20).

$$Info(D) = -\sum_{i=1}^{m} pi\ log2(pi) \quad (17)$$

$$Info_{spltpnt}(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * Info(D) \quad (18)$$

and $$Gini(D) = 1 - \sum_{i=1}^{m} pi^2 \quad (19)$$

$$Gini_{spltpnt}(D) = \frac{|D_1|}{|D|} Gini(D1) + \frac{|D_2|}{|D|} Gini(D2) \quad (20)$$

The snap shot of first 10, middle 10 and last 10 records with Information Gain & Gini Index values are shown below (Fig. 26).

We generate Info(D) and Gini(D) for every possible split-point with exception to rno 1 because it has no split point, further of 81 possible split-points 8 possible split-points do not generate info(D), this is because one of the values of fyes, fno, syes, sno is zero, as shown below (Fig. 27).

We have generated Information Gain and Gini Index for every split point; we now compare the two results.

Case 1: Information Gain

The point with minimum expected information requirement for Humidity Measured at 12:00 A.M (humid12) is to be selected as the split point; the five best cases with minimum Information Gain are shown in Fig. 28.

rno	humid12	spltpnt	fyes	fno	syes	sno	info	gini
1	18	0	0	0	0	0	0	0
2	19	18.5	0	1	1925	4025	0	0.437643
3	20	19.5	0	5	1925	4021	0	0.437502
4	21	20.5	0	7	1925	4019	0	0.437431
5	22	21.5	0	14	1925	4012	0	0.437184
6	23	22.5	0	23	1925	4003	0	0.436866
7	24	23.5	0	37	1925	3989	0	0.436369
8	25	24.5	0	47	1925	3979	0	0.436012
9	26	25.5	1	57	1924	3969	0.903642	0.435832
10	27	26.5	1	75	1924	3951	0.90198	0.435186
42	59	58.5	452	2341	1473	1685	0.828693	0.391461
43	60	59.5	483	2437	1442	1589	0.825976	0.389539
44	61	60.5	516	2543	1409	1483	0.822266	0.386989
45	62	61.5	551	2609	1374	1417	0.823413	0.387333
46	63	62.5	583	2699	1342	1327	0.820683	0.38537
47	64	63.5	600	2755	1325	1271	0.818063	0.383605
48	65	64.5	630	2817	1295	1209	0.817818	0.383169
49	66	65.5	653	2880	1272	1146	0.815476	0.381504
50	67	66.5	685	2944	1240	1082	0.815059	0.380949
51	68	67.5	711	3003	1214	1023	0.813581	0.379789
73	90	89.5	1406	3802	519	224	0.846623	0.397544
74	91	90.5	1437	3825	488	201	0.848767	0.398902
75	92	91.5	1493	3852	432	174	0.855596	0.403295
76	93	92.5	1529	3877	396	149	0.858094	0.404911
77	94	93.5	1578	3894	347	132	0.865226	0.409533
78	95	94.5	1622	3907	303	119	0.87197	0.413917
79	96	95.5	1643	3917	282	109	0.874283	0.415426
80	97	96.5	1682	3944	243	82	0.876471	0.416886
81	98	97.5	1731	3955	194	71	0.884496	0.422116
82	100	99	1741	3957	184	69	0.886153	0.423198

Fig. 26. Information Gain and Gini for each Possible Split-Point for Humidity at 12 am.

rno	humid12	spltpnt	fyes	fno	syes	sno	info	gini
2	19	18.5	0	1	1925	4025	0	0.437643
3	20	19.5	0	5	1925	4021	0	0.437502
4	21	20.5	0	7	1925	4019	0	0.437431
5	22	21.5	0	14	1925	4012	0	0.437184
6	23	22.5	0	23	1925	4003	0	0.436866
7	24	23.5	0	37	1925	3989	0	0.436369
8	25	24.5	0	47	1925	3979	0	0.436012

Fig. 27. Split-Points where Information Gain is not Generated for Humidity at 12 am.

rno	spltpnt	info
53	69.5	0.809583
54	70.5	0.810877
52	68.5	0.810984
55	71.5	0.812556
57	73.5	0.812772

Fig. 28. Five Best cases with Minimum Information Gain for Humidity at 12 am.

The above table is regenerated with Gini Index for the above split-points (Fig 29).

And in accordance to the rule of Information Gain we have to choose 69.5 as split-point for Humidity Measured at 12:00 A.M (humid12) since it has the lowest Information Gain, split-point 69.5 with all the attributes is shown below: (Fig. 30).

Case 2: Gini Index

The point giving the minimum Gini index for a given attribute Humidity Measured at 12:00 A.M (humid12) is to be taken as a split-point; the five best cases with minimum Gini Index are shown below: (Fig 31).

The above table is regenerated with Information Gain for the above split-points (Fig. 32).

And in accordance to the rule of Gini Index we have to choose 69.5 as split-point for Humidity Measured at 12:00 point 69.5 with all the attributes is shown below (Fig. 33).

rno	spltpnt	info	gini
53	69.5	0.809583	0.37666
54	70.5	0.810877	0.377227
52	68.5	0.810984	0.37783
55	71.5	0.812556	0.378023
57	73.5	0.812772	0.37766

Fig. 29. Gini Index for each Respected Split-Point.

rno	humid12	spltpnt	fyes	fno	syes	sno	info	gini
53	70	69.5	779	3151	1146	875	0.809583	0.37666

Fig. 30. Split-Point with Lowest Information Gain for Humidity at 12 am.

rno	spltpnt	gini
53	69.5	0.37666
54	70.5	0.377227
57	73.5	0.37766
52	68.5	0.37783
55	71.5	0.378023

Fig. 31. Five Best Cases with Minimum Gini Index for Humidity at 12 am.

rno	spltpnt	gini	info
53	69.5	0.37666	0.809583
54	70.5	0.377227	0.810877
57	73.5	0.37766	0.812772
52	68.5	0.37783	0.810984
55	71.5	0.378023	0.812556

Fig. 32. Information Gain for each Respected Split-Point.

rno	humid12	spltpnt	fyes	fno	syes	sno	info	gini
53	70	69.5	779	3151	1146	875	0.809583	0.37666

Fig. 33. Split-Point with Lowest Gini Index for Humidity at 12 am.

The results of Information Gain and Gini Index are exactly the same, hence split-point 69.5 will be chosen in either case, and there is no conflict at all.

5) Humidity Measured at 03:00 P.M: Like the earlier three cases Humidity Measured at 03:00 P.M (humid3) is also continuous valued rather than discrete valued, and accordingly best split- point for humidity3 is generated and evaluated as well.

In the dataset of 5951 records there are 80 unique values of Humidity Measured at 03:00 P.M (humid3) recoded, varying from 16 to 100. The snap shot of first 10, middle 10 and last 10-sorted records with mid points are shown below (Fig 34), 79 possible split-points will be evaluated.

The snap shot of first 10, middle 10 and last 10 records with necessary values of fyes, fno, syes & sno are shown in Fig 35.

rno	humid3	spltpnt
1	16	0
2	17	16.5
3	20	18.5
4	22	21
5	24	23
6	25	24.5
7	26	25.5
8	27	26.5
9	28	27.5
10	30	29
41	61	60.5
42	62	61.5
43	63	62.5
44	64	63.5
45	65	64.5
46	66	65.5
47	67	66.5
48	68	67.5
49	69	68.5
50	70	69.5
71	91	90.5
72	92	91.5
73	93	92.5
74	94	93.5
75	95	94.5
76	96	95.5
77	97	96.5
78	98	97.5
79	99	98.5
80	100	99.5

Fig. 34. Unique Values of Humidity at 3 pm and their Split Points.

rno	humid3	spltpnt	fyes	fno	syes	sno
1	16	0	0	0	0	0
2	17	16.5	0	2	1925	4024
3	20	18.5	0	3	1925	4023
4	22	21	0	6	1925	4020
5	24	23	0	7	1925	4019
6	25	24.5	0	9	1925	4017
7	26	25.5	0	10	1925	4016
8	27	26.5	0	11	1925	4015
9	28	27.5	0	14	1925	4012
10	30	29	0	15	1925	4011
41	61	60.5	54	752	1871	3274
42	62	61.5	64	813	1861	3213
43	63	62.5	76	890	1849	3136
44	64	63.5	93	953	1832	3073
45	65	64.5	113	1036	1812	2990
46	66	65.5	126	1128	1799	2898
47	67	66.5	147	1222	1778	2804
48	68	67.5	168	1314	1757	2712
49	69	68.5	177	1420	1748	2606
50	70	69.5	211	1523	1714	2503
71	91	90.5	1101	3522	824	504
72	92	91.5	1193	3593	732	433
73	93	92.5	1272	3690	653	336
74	94	93.5	1371	3767	554	259
75	95	94.5	1468	3817	457	209
76	96	95.5	1522	3853	403	173
77	97	96.5	1603	3929	322	97
78	98	97.5	1710	3946	215	80
79	99	98.5	1730	3949	195	77
80	100	99.5	1731	3949	194	77

Fig. 35. Possible Splitpoints for Humidity at 3 pm.

For each possible split-point for Minimum Temperature, we will calculate $Info_{spltpnt}(D)$ and $Gini_{spltpnt}(D)$ using following equations (21)(22)(23)(24).

$$Info(D) = -\sum_{i=1}^{m} pi\ log2(pi) \quad (21)$$

$$Info_{spltpnt}(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * Info(D) \quad (22)$$

and $$Gini(D) = 1 - \sum_{i=1}^{m} pi^2 \quad (23)$$

$$Gini_{spltpnt}(D) = \frac{|D_1|}{|D|} Gini(D1) + \frac{|D_2|}{|D|} Gini(D2) \quad (24)$$

The snap shot of first 10, middle 10 and last 10 records with Information Gain & Gini Index values are shown below: (Fig. 36).

We generate Info(D) and Gini(D) for every possible split-point with exception to rno 1 because it has no split point, further of 79 possible split-points 15 possible split-points do not generate info(D), this is because one of the values of fyes, fno, syes, sno is zero, as shown below: (Fig. 37).

We have generated Information Gain and Gini Index for every split point; we now compare the two results.

Case 1: Information Gain

The point with minimum expected information requirement for Humidity Measured at 03:00 P.M (humid3) is to be selected as the split point for Humidity Measured at 03:00 P.M (humid3) the five best cases with minimum Information Gain are shown below: (Fig 38).

rno	humid3	spltpnt	fyes	fno	syes	sno	info	gini
1	16	0	0	0	0	0	0	0
2	17	16.5	0	2	1925	4024	0	0.437608
3	20	18.5	0	3	1925	4023	0	0.437572
4	22	21	0	6	1925	4020	0	0.437467
5	24	23	0	7	1925	4019	0	0.437431
6	25	24.5	0	9	1925	4017	0	0.437361
7	26	25.5	0	10	1925	4016	0	0.437326
8	27	26.5	0	11	1925	4015	0	0.43729
9	28	27.5	0	14	1925	4012	0	0.437184
10	30	29	0	15	1925	4011	0	0.437149
41	61	60.5	54	752	1871	3274	0.865622	0.417068
42	62	61.5	64	813	1861	3213	0.863979	0.415986
43	63	62.5	76	890	1849	3136	0.861462	0.414452
44	64	63.5	93	953	1832	3073	0.861832	0.414211
45	65	64.5	113	1036	1812	2990	0.86107	0.413424
46	66	65.5	126	1128	1799	2898	0.856965	0.411126
47	67	66.5	147	1222	1778	2804	0.855041	0.409773
48	68	67.5	168	1314	1757	2712	0.853037	0.408397
49	69	68.5	177	1420	1748	2606	0.845842	0.404508
50	70	69.5	211	1523	1714	2503	0.84627	0.404191
71	91	90.5	1101	3522	824	504	0.828948	0.386998
72	92	91.5	1193	3593	732	433	0.837883	0.392434
73	93	92.5	1272	3690	653	336	0.838367	0.392463
74	94	93.5	1371	3767	554	259	0.845895	0.397129
75	95	94.5	1468	3817	457	209	0.857442	0.40452
76	96	95.5	1522	3853	403	173	0.861853	0.407349
77	97	96.5	1603	3929	322	97	0.86224	0.407678
78	98	97.5	1710	3946	215	80	0.882094	0.420539
79	99	98.5	1730	3949	195	77	0.885633	0.42285
80	100	99.5	1731	3949	194	77	0.88584	0.422986

Fig. 36. Information Gain and Gini for each Possible Split-Point for Humidity at 3 pm.

rno	humid3	spltpnt	fyes	fno	syes	sno	info	gini
2	17	16.5	0	2	1925	4024	0	0.437608
3	20	18.5	0	3	1925	4023	0	0.437572
4	22	21	0	6	1925	4020	0	0.437467
5	24	23	0	7	1925	4019	0	0.437431
6	25	24.5	0	9	1925	4017	0	0.437361
7	26	25.5	0	10	1925	4016	0	0.437326
8	27	26.5	0	11	1925	4015	0	0.43729
9	28	27.5	0	14	1925	4012	0	0.437184
10	30	29	0	15	1925	4011	0	0.437149
11	31	30.5	0	20	1925	4006	0	0.436972
12	32	31.5	0	22	1925	4004	0	0.436901
13	33	32.5	0	25	1925	4001	0	0.436795
14	34	33.5	0	28	1925	3998	0	0.436689
15	35	34.5	0	40	1925	3986	0	0.436262
16	36	35.5	0	53	1925	3973	0	0.435797

Fig. 37. Split-Points where Information Gain is not Generated for Humidity at 3 pm.

rno	spltpnt	info
63	82.5	0.817457
70	89.5	0.818059
67	86.5	0.819001
64	83.5	0.819493
68	87.5	0.819696

Fig. 38. Five Best cases with Minimum Information Gain for Humidity at 3 pm.

The above table is regenerated with Gini Index for the above split-points (Fig. 39).

And in accordance to the rule of Information Gain we have to choose 82.5 as split-point for Humidity Measured at 03:00 P.M (humid3) since it has the lowest Information Gain, split-point 82.5 with all the attributes is shown in Fig. 40.

rno	spltpnt	info	gini
63	82.5	0.817457	0.38305
70	89.5	0.818059	0.380388
67	86.5	0.819001	0.382286
64	83.5	0.819493	0.38386
68	87.5	0.819696	0.382274

Fig. 39. Gini Index for each Respected Split-Point.

rno	humid3	spltpnt	fyes	fno	syes	sno	info	gini
63	83	82.5	619	2798	1306	1228	0.817457	0.38305

Fig. 40. Split-Point with Lowest Information Gain for Humidity at 3 pm.

Case 2: Gini Index

The point giving the minimum Gini index for a given attribute Humidity Measured at 03:00 P.M (humid3) is to be taken as a split-point for the Humidity Measured at 03:00 P.M (humid3) the five best cases with minimum Gini Index are shown below: (Fig. 41).

The above table is regenerated with Information Gain for the above split-points (Fig. 42).

And in accordance to the rule of Gini Index we have to choose 89.5 as split-point for Humidity Measured 03:00 P.M (humid3) since it has the lowest Gini Index, split-point 89.5 with all the attributes is shown below (Fig. 43).

As per Information Gain choice of split-point is 82.5, while as per the choice of Gini Index the split-point is 89.5. In order to make decision on the choice of split-point we compare the two generated list, as shown below (Fig. 44).

rno	spltpnt	gini
70	89.5	0.380388
68	87.5	0.382274
67	86.5	0.382286
69	88.5	0.382833
63	82.5	0.38305

Fig. 41. Five best cases with Minimum Gini Index for Humidity at 3 pm

rno	spltpnt	gini	info
70	89.5	0.380388	0.818059
68	87.5	0.382274	0.819696
67	86.5	0.382286	0.819001
69	88.5	0.382833	0.821271
63	82.5	0.38305	0.817457

Fig. 42. Information Gain for each Respected Split-Point.

rno	humid3	spltpnt	fyes	fno	syes	sno	info	gini
70	90	89.5	1013	3463	912	563	0.818059	0.380388

Fig. 43. Split-Point with Lowest Gini Index for Humidity at 3 pm.

Information Gain VS Gini Index

rno	spltpnt	info
63	82.5	0.817457
70	89.5	0.818059
67	86.5	0.819001
64	83.5	0.819493
68	87.5	0.819696

rno	spltpnt	gini
70	89.5	0.380388
68	87.5	0.382274
67	86.5	0.382286
69	88.5	0.382833
63	82.5	0.38305

Fig. 44. Comparison between Information Gain and Gini Index.

From the comparison shown above, there is a visible overlap between the two results, we choose 89.5 as split-point for Humidity Measured at 03:00 P.M (humid3), because it is first choice as per Gini Index and it is second choice of Information Gain.

B. Evaluation -- Information Gain vs. Gini Index

Four attributes are continuous valued rather than discrete valued, we employed Information Gain used by ID3 and Gini Index used by CART to determine best possible split-point, the results are shown below (Table I).

TABLE I. BEST POSSIBLE SPLITS USING ID3 AND CART

Attribute	Information Gain	Gini Index	Class One	Class Two
TMAX	25.05	8.05	8.05<= is H1	>8.05 is H2
TMIN	-0.35	-0.35	-0.35<= is L1	>-0.35 is L2
HUMID12	69.5	69.5	69.5<= is T1	>69.5 is T2
HUMID3	82.5	89.5	89.5<= is U1	>89.5 is U2

Of the four attributes, Tmin and Humid12 have same results for Information Gain and Gini Index. Humid3 has overlapping results for Information Gain and Gini Index, as already discussed we choose 89.5 as split-point for Humid3. It is the attribute Tmax where the results of Information Gain and Gini Index do not corroborate, and hence we have to choose one of the values, either as per Information Gain (25.05) or as per Gini Index (8.05). We chose Gini Index over Information Gain primarily because the split-point of three attributes (Tmin, Humid12, Humid3) is as per Gini Index while as split point of two attributes (Tmin & Humid12) is as per Information Gain, thus we choose to go with the majority i.e. Gini Index over Information Gain accordingly split-point of Tmax is 8.05.

C. Rest of Data Attributes

Off the rest of the data attributes, Station_id, Year, Month and date, we decide not to consider recording station (Station_id) as part of decision tree for prediction of rainfall, since all the stations belong to the same province. Further, a year is 365 days or 12 month or 4 seasons, thus we split the months into season as shown below: (Table II).

TABLE II. SPLITTING MONTHS IN RESPECTED SEASONS

Months	Season
12, 1, 2	Winter
3, 4, 5	Spring
6, 7, 8	Summer
9, 10, 11	Autumn

Thus we use seasons instead of months, and decide not to use year and date as part of decision table, this will also maximize information dissemination.

1) Resultant Dataset: Consequent upon conversion of continuous valued attributes into discrete valued and conversion of months into seasons besides not considering some irrelevant attributes, the snapshot of the resultant dataset is shown below: (Fig. 45).

season	ctmax	ctmin	chumid12	chumid3	crfall
spring	H1	L1	T2	U1	N
spring	H1	L1	T2	U2	Y
spring	H2	L1	T1	U1	N
spring	H2	L2	T1	U1	N
spring	H2	L2	T1	U1	N
spring	H2	L2	T2	U1	Y
spring	H2	L2	T2	U2	Y
spring	H2	L2	T2	U1	Y
spring	H2	L2	T2	U2	Y
spring	H1	L2	T2	U2	Y
spring	H2	L2	T2	U1	Y
spring	H2	L2	T2	U2	Y
spring	H1	L2	T1	U1	Y
spring	H2	L1	T2	U1	Y
spring	H2	L2	T2	U2	Y
spring	H1	L1	T2	U2	Y
spring	H1	L1	T2	U1	Y
spring	H1	L1	T2	U2	Y
spring	H2	L1	T1	U1	N
spring	H2	L2	T1	U1	N

Fig. 45. Labelled Resultant Dataset.

Where

Ctmax = H1 if tmax <= **8.05**

Ctmax = H2 if tmax > **8.05**

Ctmin = L1 if tmin <= **-0.35**

Ctmin = L2 if tmin > **-0.35**

Chumid12 = T1 if humid12 <= **69.5**
Chumid12 = T2 if humid12 > **69.5**

Chumid3 = U1 if humid3 <= **89.5**
Chumid3 = U2 if humid3 > **89.5**

Further months have been converted into seasons as per the table shown above and crfall is Y if rfall >0 and crfall is N if rfall =0.

IV. CONCLUSION AND FUTURE WORK

In this paper two techniques are employed i.e. Information Gain and Gini index to convert continuous data into discrete valued data. This is preliminary and prerequisite step in order to apply machine learning algorithm Decision tree on the geographical data set. Besides having prepared historical geographical data for the application of Decision tree algorithm we have also compared the results from two varying techniques applied on the same dataset.

Whilst this study was primarily aimed at the comparison of Information Gain and Gini index, a fuller work is underway in which two separate dataset shall be generated on the basis of Information Gain and Gini index thereafter decision tree

algorithms shall be employed on these two generated data sets this will enable us to compare the performance of Information Gain and Gini index at the individual level of implementation.

REFERENCES

[1] Han, J., Kamber, M.: Data Mining Concepts and Techniques. China Machine Press, Beijing (2007).

[2] Zhang Quancheng, You Kun, Ma, Gang. Application of ID3 Algorithm in Exercise Prescription[C]. The International Conference on Electric and Electronics, Nanchang, China, June 22, 2011. 99(3): 669-675.

[3] LI Shoubang. Application Study on Mining of University Students ' Physical Fitness Test Data Based on Classification Rules: Taking the Juniors of Xi'an Shiyou University as an Example [J]. Journal of Xi'an Shiyou University (Natural Science Edition), 2018, 33(5) : 120- 126.

[4] J.R. Quinlan. Improved use of continuous attributes in c4.5. Journal of Artificial Intelligence Approach, (4):77–90, 1996. 211, 212, 216.

[5] Data Mining: Concepts and Techniques, 3rd Edition Jiawei Han, Micheline Kamber, Jian Pei.

[6] Ashraf, Mudasir, Majid Zaman, and Muheet Ahmed. "Performance analysis and different subject combinations: An empirical and analytical discourse of educational data mining." 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2018.

[7] Ashraf, Mudasir, Majid Zaman, and Muheet Ahmed. "Using Ensemble StackingC Method and Base Classifiers to Ameliorate Prediction Accuracy of Pedagogical Data." Procedia computer science 132 (2018): 1021-1040.

[8] Mirza, Shuja, Sonu Mittal, and Majid Zaman. "A Review of Data Mining Literature." International Journal of Computer Science and Information Security (IJCSIS) 14.11 (2016).

[9] Shuja, Mirza, Sonu Mittal, and Majid Zaman. "Diabetes Mellitus and Data Mining Techniques: A survey.". International Journal of Computer Sciences and Engineering 7 (2019): 858.

[10] Mirza, Shuja, Sonu Mittal, and Majid Zaman. "Decision Support Predictive model for prognosis of diabetes using SMOTE and Decision tree." International Journal of Applied Engineering Research 13.11 (2018): 9277-9282.

[11] Shuja M., Mittal S., & Zaman M. (2018). Decision Support System for Prognosis of Diabetes using Non-Clinical Parameters and Data Mining Techniques, International Journal of Database Theory and Applications, 11(3), 39-48.

[12] Shuja, Mirza, Sonu Mittal, and Majid Zaman. "Effective Prediction of Type II Diabetes Mellitus Using Data Mining Classifiers and SMOTE." Advances in Computing and Intelligent Systems. Springer, Singapore, 2020. 195-211.

[13] Raileanu, Laura Elena, and Kilian Stoffel. "Theoretical comparison between the Gini index and Information gain criteria." Annals of Mathematics and Artificial Intelligence 41.1 (2004): 77-93.

[14] Muharram M.A., Smith G.D. (2004) Evolutionary Feature Construction Using Information Gain and Gini Index. In: Keijzer M., O'Reilly UM., Lucas S., Costa E., Soule T. (eds) Genetic Programming. EuroGP 2004. Lecture Notes in Computer Science, vol 3003. Springer, Berlin, Heidelberg.

[15] Kulkarni, Vrushali Y., Manisha Petare, and P. K. Sinha. "Analyzing random forest classifier with different split measures." Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012. Springer, New Delhi, 2014.