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

Dr. Majid Zaman¹
Directorate of IT & SS
University of Kashmir
Srinagar, India

Sameer Kaul²
Department of Computer Science, University of Kashmir
Srinagar, India

Dr. Muheet Ahmed³
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).

\[
\text{Info}(D) = -\sum_{i=1}^{m} \pi_i \log_2(\pi_i) [5] 
\]

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

\[
\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D) \quad [5]
\]

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].

\[
\text{Gini}(D) = 1 - \sum_{i=1}^{m} \pi_i^2 \quad [4]
\]

\[
\text{Gini}(D) = \frac{|D_1|}{|D|} \text{Gini}(D_1) + \frac{|D_2|}{|D|} \text{Gini}(D_2) \quad [5]
\]

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

\[
\text{Gini}(A) = \text{Gini}(D) - \text{Gini}(A) \quad [6]
\]

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].

www.ijacsa.thesai.org

429 | P a g e
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.
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 1925 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.

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\(\text{max}(D)\) and Gini\(\text{max}(D)\) 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) fyes: 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\(\text{splitpt}(D)\) and Gini\(\text{splitpt}(D)\) using following equations

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

\[
\text{Info}_{\text{splitpt}}(D) = \sum_{j=1}^{v} \left| \frac{D_j}{|D|} \right| \times \text{Info}(D) \tag{9}\]

\[
\text{and Gini}(D) = 1 - \sum_{i=1}^{m} pi^2 \tag{10}\]

\[
\text{Gini}_{\text{splitpt}}(D) = \left| \frac{D_1}{|D|} \right| \text{Gini}(D1) + \left| \frac{D_2}{|D|} \right| \text{Gini}(D2) \tag{11}\]

www.ijacsa.thesai.org
For example, we will generate $\text{Info}(D)$ for a possible split-point of 10.85 listed in above table for rno 144 for $t_{\text{max}}$ of 10.9.

\[
\text{Info}_{\text{spltpnt}}(D) = \sum_{j=1}^{N} \left| D_j \right| * \text{Info}(D) \tag{12}
\]

\[
\text{Info}_{10.9}(D) = \frac{1552}{5951} * \left( -\frac{665}{1552} \times \text{LOG2} \left( \frac{665}{1552} \right) \right) + \frac{4399}{5951} * \left( -\frac{1260}{4399} \times \text{LOG2} \left( \frac{1260}{4399} \right) \right) - \frac{3139}{4399} * \left( -\frac{3139}{4399} \right) \tag{10}
\]

\[
\text{Info}_{10.9}(D) = 0.895652633
\]

And we will generate $\text{Gini}(D)$ for a possible split-point of 10.85 listed in above table for rno 144 for $t_{\text{max}}$ of 10.9.

\[
\text{Gini}_{\text{spltpnt}}(D) = \left[ \frac{|D_1|}{|D|} \times \text{Gini}(D_1) \right] + \left[ \frac{|D_2|}{|D|} \times \text{Gini}(D_2) \right] \tag{5}
\]

\[
\text{Gini}_{10.9}(D) = \left[ \frac{1552}{5951} \times \left( 1 - \left( \frac{665}{1552} \right)^2 - \frac{887}{1552} \right) \right] + \frac{4399}{5951} * \left( 1 - \left( \frac{1260}{4399} \right)^2 - \frac{3139}{4399} \right) \tag{11}
\]

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

Likewise we generate $\text{Info}(D)$ and $\text{Gini}(D)$ 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 $\text{Info}(D)$ and $\text{Gini}(D)$ 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 $\text{Info}(D)$, show below (Fig. 7).

This is because one of the values of $f_{\text{yes}}$, $f_{\text{no}}$, $s_{\text{yes}}$, $s_{\text{no}}$ 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 ($t_{\text{max}}$) is to be selected as the split point for Maximum Temperature ($t_{\text{max}}$), 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 | $t_{\text{max}}$ | spltpnt | $f_{\text{yes}}$ | $f_{\text{no}}$ | $s_{\text{yes}}$ | $s_{\text{no}}$ | info | gini |
|-----|----------------|---------|-----------------|----------------|----------------|----------------|------|-----|
| 1   | -8.2           | 0       | 0               | 0              | 0              | 0              | 0    | 0   |
| 2   | -7.8           | -8      | 0               | 1              | 925            | 4025           | 0    | 0.4797643 |
| 3   | -7.6           | -7.7    | 0               | 2              | 1925           | 4024           | 0    | 0.4760804 |
| 4   | -7.5           | -7.5    | 0               | 3              | 1925           | 4023           | 0    | 0.4757023 |
| 5   | -6.7           | -7.1    | 0               | 5              | 1925           | 4021           | 0    | 0.4750892 |
| 6   | -5.6           | -6.1    | 0               | 6              | 1925           | 4020           | 0    | 0.4748738 |
| 7   | -5             | -5.25   | 1               | 7              | 1924           | 4019           | 0    | 0.4748618 |
| 8   | -4.6           | -4.75   | 2               | 7              | 1923           | 4019           | 0    | 0.4748618 |
| 9   | -4.4           | -4.45   | 3               | 7              | 1922           | 4019           | 0    | 0.4748618 |
| 10  | -4             | -4.2    | 3               | 8              | 1922           | 4018           | 0    | 0.4748618 |
| 124 | 19.6           | 19.65   | 667             | 968            | 1278           | 3168           | 0.8986450 | 0.43081 |
| 142 | 19.7           | 10.65   | 655             | 972            | 1272           | 3154           | 0.8986450 | 0.43081 |
| 143 | 19.6           | 10.75   | 655             | 976            | 1272           | 3154           | 0.8986450 | 0.43081 |
| 144 | 19.9           | 10.85   | 666             | 987            | 1260           | 3139           | 0.8986450 | 0.43081 |
| 145 | 11             | 10.95   | 696             | 986            | 1259           | 3130           | 0.8986450 | 0.43081 |
| 146 | 11             | 11.05   | 674             | 972            | 1251           | 3134           | 0.8986450 | 0.43081 |
| 147 | 11.2           | 11.15   | 674             | 933            | 1251           | 3134           | 0.8986450 | 0.43081 |
| 148 | 11.3           | 11.25   | 675             | 951            | 1247           | 3075           | 0.8986450 | 0.43081 |
| 149 | 11.4           | 11.35   | 678             | 965            | 1247           | 3061           | 0.8986450 | 0.43081 |
| 150 | 11.5           | 11.45   | 681             | 963            | 1244           | 3043           | 0.8986450 | 0.43081 |
| 371 | 33.6           | 33.55   | 1922           | 4085           | 3              | 21             | 0     | 0   |
| 372 | 33.7           | 33.65   | 1923           | 4088           | 2              | 18             | 0     | 0   |
| 373 | 33.6           | 33.75   | 1923           | 4089           | 2              | 17             | 0     | 0   |
| 374 | 33.9           | 33.85   | 1923           | 4083           | 2              | 13             | 0     | 0   |
| 375 | 36.4           | 33.95   | 1923           | 4082           | 2              | 9              | 0     | 0   |
| 376 | 34.1           | 34.05   | 1924           | 4019           | 1              | 7              | 0     | 0   |
| 377 | 34.2           | 34.15   | 1925           | 4021           | 0              | 5              | 0     | 0   |
| 378 | 34.4           | 34.3    | 1926           | 4022           | 0              | 2              | 0     | 0   |
| 379 | 34.6           | 34.5    | 1926           | 4024           | 0              | 2              | 0     | 0   |
| 380 | 35.4           | 35      | 1925           | 4025           | 0              | 1              | 0     | 0   |

Fig. 5. Possible Splitpoints for Maximum Temperature.

In this way we generate $\text{Info}(D)$ and $\text{Gini}(D)$ 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).
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).

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 midpoint between each pair of adjacent values is generated as possible split-point, the snapshot 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.
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 $\text{Info}_{\text{splitpt}}(D)$ and $\text{Gini}_{\text{splitpt}}(D)$ using following equations. (13)(14)(15)(16).

\[
\text{Info}(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{13}
\]

\[
\text{Info}_{\text{splitpt}}(D) = \sum_{j=1}^{n} \frac{|D_j|}{|D|} \times \text{Info}(D) \tag{14}
\]

and \[ Gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \tag{15}\]

\[
\text{Gini}_{\text{splitpt}}(D) = \frac{|D_1|}{|D|} \times \text{Gini}(D_1) + \frac{|D_2|}{|D|} \times \text{Gini}(D_2) \tag{16}\]

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 row 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.

![Fig. 15. Possible Splitpoints for Minimum Temperature.](image-url)

![Fig. 16. Information Gain and Gini for each Possible Split-Point for Minimum Temperature.](image-url)

![Fig. 17. Split-Points where Information Gain is not Generated for Minimum Temperature.](image-url)

**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).

![Fig. 18. Five Best cases with Minimum Information Gain for Minimum Temperature.](image-url)
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.

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

\[
\text{Info}(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) 
\]

(17)

\[
\text{Info}_{\text{spltpnt}}(D) = \sum_{i=1}^{m} \frac{|p_i|}{|p|} \times \text{Info}(D)
\]

(18)

\[
\text{Gini}(D) = 1 - \sum_{i=1}^{m} p_i^2
\]

(19)

\[
\text{Gini}_{\text{spltpnt}}(D) = \frac{|p_1|}{|p|} \text{Gini}(D1) + \frac{|p_2|}{|p|} \text{Gini}(D2)
\]

(20)

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

We generate $\text{Info}(D)$ and $\text{Gini}(D)$ 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, sys, 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.

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

**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).
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) recorded, 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.

For each possible split-point for Minimum Temperature, we will calculate $\text{Info}_{\text{splitpt}}(D)$ and $\text{Gini}_{\text{splitpt}}(D)$ using following equations (21)(22)(23)(24).

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

And in accordance to the rule of Gini Index we have to choose 89.5 as split-point for Humidity Measured at 03:00 P.M (humid3) since it has the lowest Gini Index, split-point 89.5 with all the attributes is shown below (Fig. 40).

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. 41).

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

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

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

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 accordingly split-point of TMAX is 8.05.
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(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 (IJCSI) 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.