Title : Prediction of the "play" attribute in the weather dataset

Group Details :

Aniruddha Hore	1PI13IS019
Giri Gaurav Bhatnagar	1PI13IS039
Kishan Kishore	1PI13IS051

Guide : Dr.Shylaja S S
Prof. And Head
Dept. Of ISE
PESIT

Introduction

This project tries to implement a machine learning program in Java using the Weka library.

Using J48 classifier , we try to predict the label (value) associated with an attribute ('play') of an instance from the given dataset ('weather').

Objective

The weather dataset contains many attributes including 'play' along with other attributes related to weather of a given day.

Our objective is to train a classifier from the 'weather dataset' and apply it to a 'test dataset' and then predict whether the attribute play bears the value 'YES' or 'NO'

Motivation

Machine learning is a vast and a crucial technology that is being used for the better decision making and smart prediction .

Feasibility

- Collection and maintenance of a big data set is a little difficult.

- Prediction will be more accurate if a large dataset is involved .

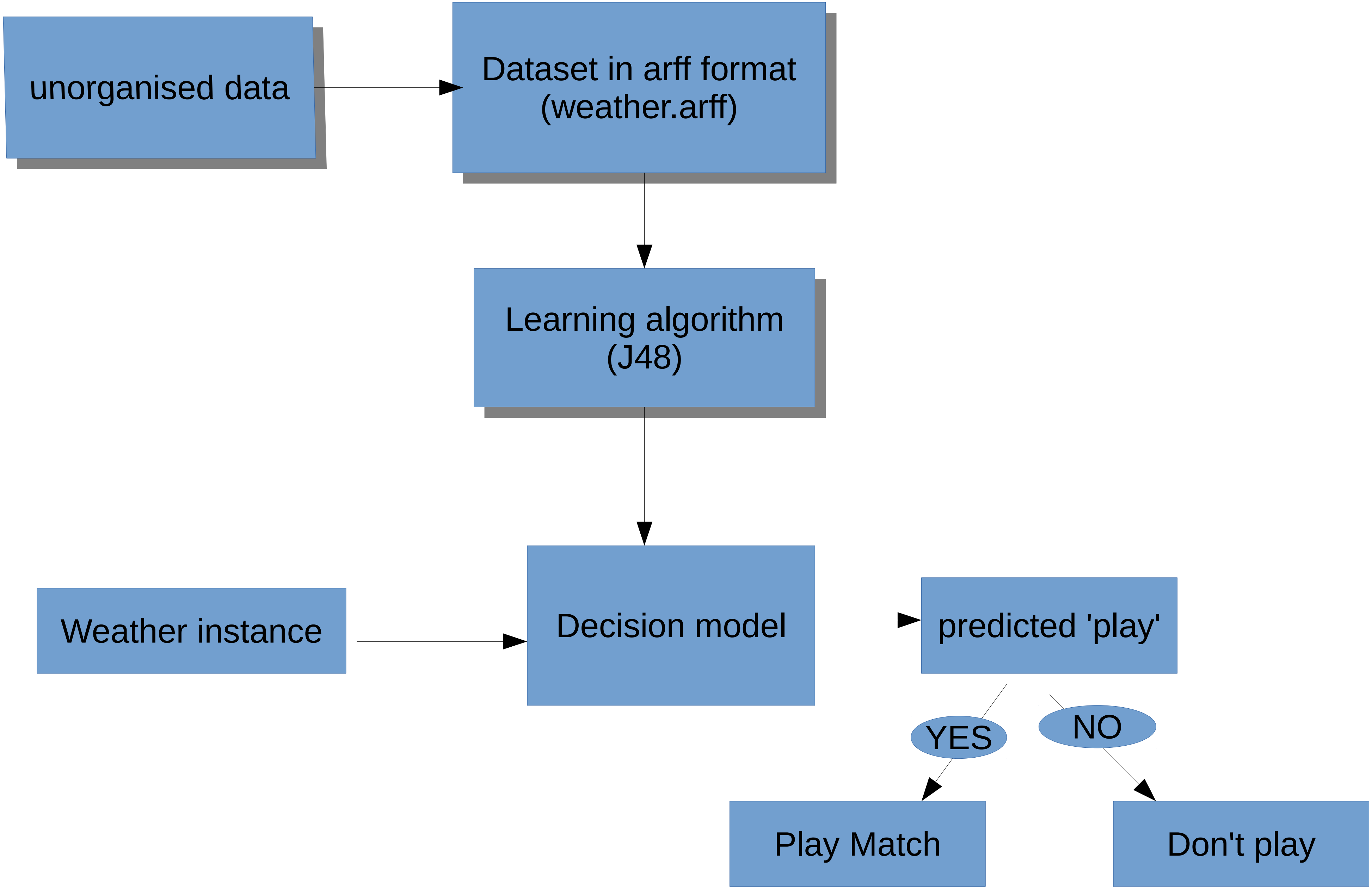
Tools/Software/Languages

Java

Weka 3

weatherunderground

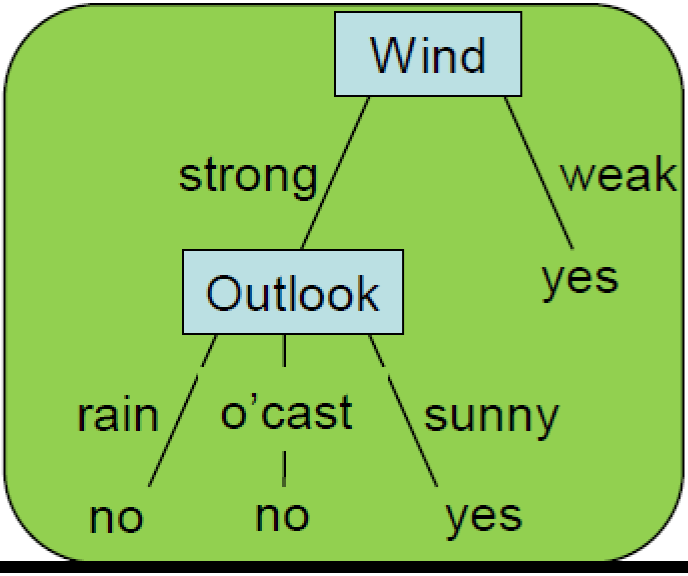
SYSTEM DESIGN

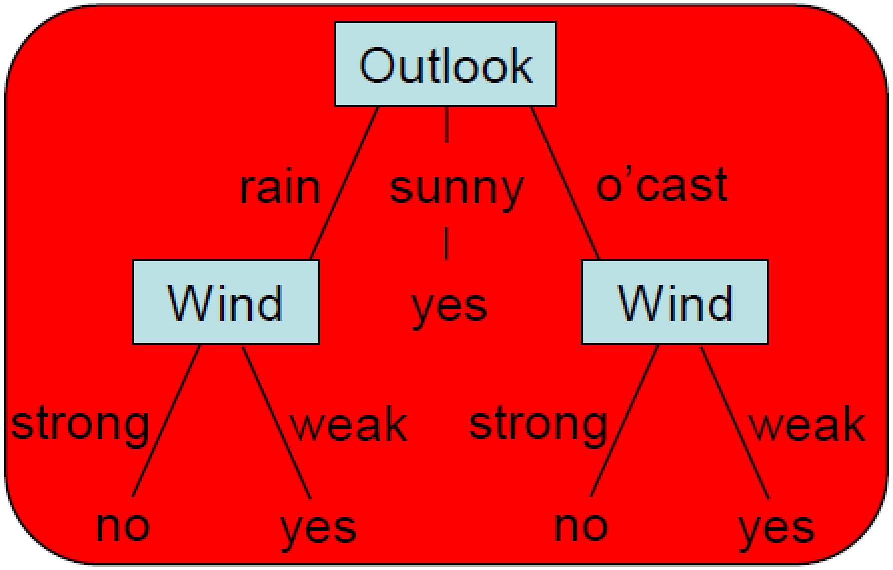


How do we produce decision trees?

- Basic Algorithm is recursive…
  - Determine which attribute to use as the top most node of the decision tree (candidates being “Wind” or “Outlook” in the first step using this example)
    - do this by calculating the information gain for each candidate attribute
      - pick the attribute with the highest information gain.

Outlook	Wind	PlayGolf
rain	strong	no
sunny	weak	yes
overcast	weak	yes
rain	weak	yes
sunny	strong	yes
rain	strong	no
overcast	strong	no





Calculating Information Gain

- Determine which attribute to use as the top most node of the decision tree – do this by calculating the information gain for each candidate attribute – pick the attribute with the highest information gain.

$$\mathbf{Gain(S, A) \equiv Entropy(S)} \; - \sum_{v\, \varepsilon Values(A)} (|S_v| \,/\, |S|)\mathbf{Entropy(S_v)}$$

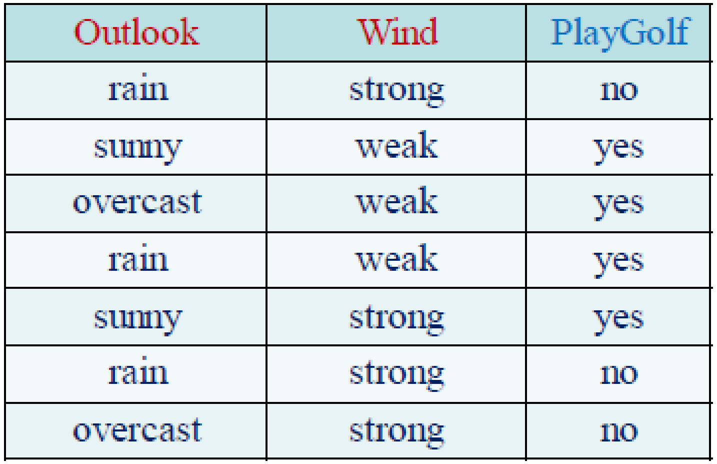
The information Gain of attribute **A** in collection **S** where **Values(A)** is the set of possible values for attribute **A** and **S_v** is the subset of **S** for which attribute **A** has the value **v**

$$\mathbf{Entropy(S)} \equiv -p_+\log_2 p_+ \; - \; p_-\log_2 p_-$$

Where target classification is boolean

p_+ :the proportion of positive examples in collection S

p_- :the proportion of negative examples in collection S



Outlook	Wind	PlayGolf
rain	strong	no
sunny	weak	yes
overcast	weak	yes
rain	weak	yes
sunny	strong	yes
rain	strong	no
overcast	strong	no

Modules

- prepareDataset

  // prepares the arff file from a csv file

- evaluate

  // evaluates the model , used to calculate acurracy

- readfile

  // reads a datafile

- addInstances

  // add your own instances

- display

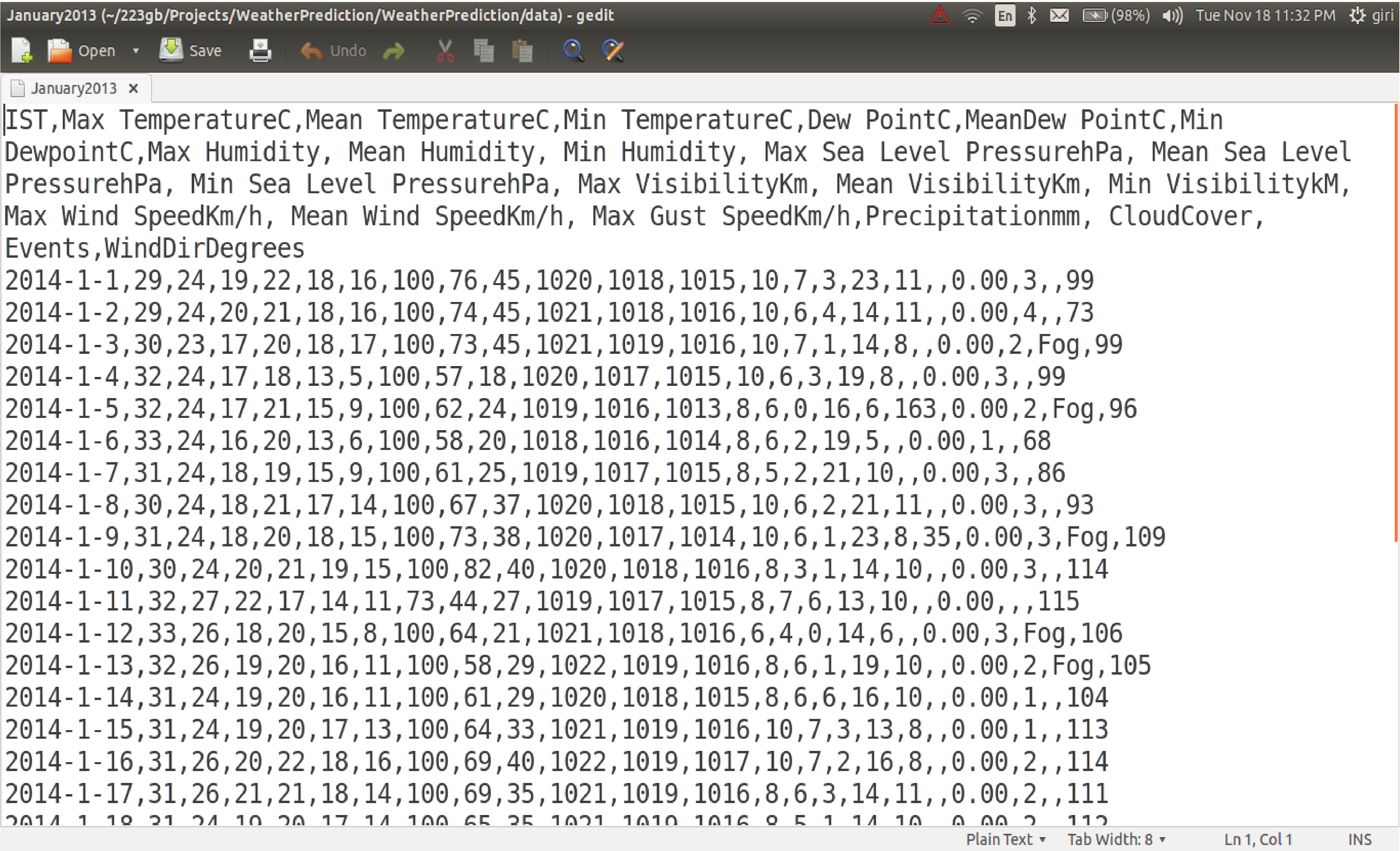
  //display the tree

- crossValidation

  //prepares dataset pairs for crossvalidation

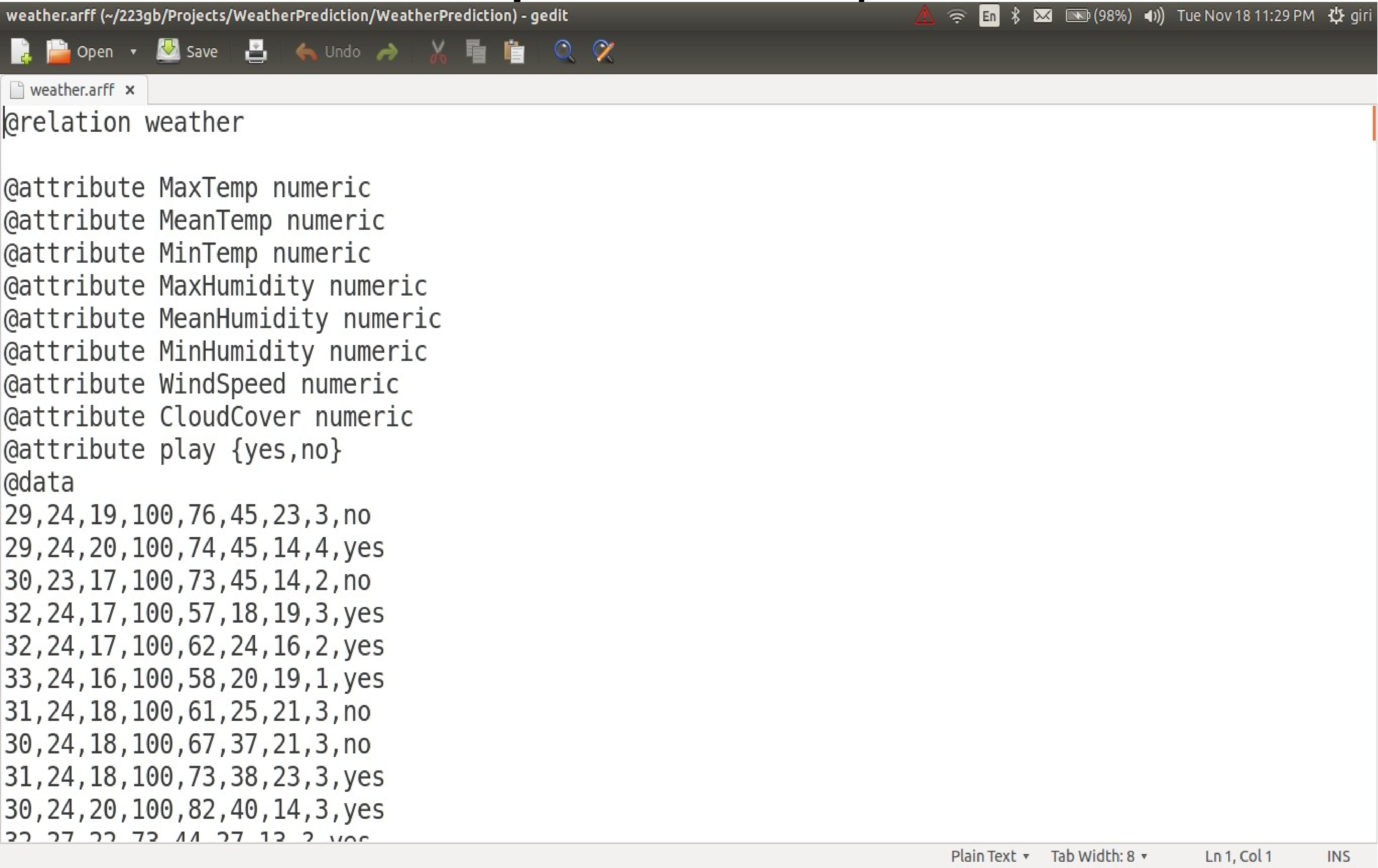
Sample Dataset

[csv format]



```
IST,Max TemperatureC,Mean TemperatureC,Min TemperatureC,Dew PointC,MeanDew PointC,Min
DewpointC,Max Humidity, Mean Humidity, Min Humidity, Max Sea Level PressurehPa, Mean Sea Level
PressurehPa, Min Sea Level PressurehPa, Max VisibilityKm, Mean VisibilityKm, Min VisibilitykM,
Max Wind SpeedKm/h, Mean Wind SpeedKm/h, Max Gust SpeedKm/h,Precipitationmm, CloudCover,
Events,WindDirDegrees
2014-1-1,29,24,19,22,18,16,100,76,45,1020,1018,1015,10,7,3,23,11,,0.00,3,,99
2014-1-2,29,24,20,21,18,16,100,74,45,1021,1018,1016,10,6,4,14,11,,0.00,4,,73
2014-1-3,30,23,17,20,18,17,100,73,45,1021,1019,1016,10,7,1,14,8,,0.00,2,Fog,99
2014-1-4,32,24,17,18,13,5,100,57,18,1020,1017,1015,10,6,3,19,8,,0.00,3,,99
2014-1-5,32,24,17,21,15,9,100,62,24,1019,1016,1013,8,6,0,16,6,163,0.00,2,Fog,96
2014-1-6,33,24,16,20,13,6,100,58,20,1018,1016,1014,8,6,2,19,5,,0.00,1,,68
2014-1-7,31,24,18,19,15,9,100,61,25,1019,1017,1015,8,5,2,21,10,,0.00,3,,86
2014-1-8,30,24,18,21,17,14,100,67,37,1020,1018,1015,10,6,2,21,11,,0.00,3,,93
2014-1-9,31,24,18,20,18,15,100,73,38,1020,1017,1014,10,6,1,23,8,35,0.00,3,Fog,109
2014-1-10,30,24,20,21,19,15,100,82,40,1020,1018,1016,8,3,1,14,10,,0.00,3,,114
2014-1-11,32,27,22,17,14,11,73,44,27,1019,1017,1015,8,7,6,13,10,,0.00,,,115
2014-1-12,33,26,18,20,15,8,100,64,21,1021,1018,1016,6,4,0,14,6,,0.00,3,Fog,106
2014-1-13,32,26,19,20,16,11,100,58,29,1022,1019,1016,8,6,1,19,10,,0.00,2,Fog,105
2014-1-14,31,24,19,20,16,11,100,61,29,1020,1018,1015,8,6,6,16,10,,0.00,1,,104
2014-1-15,31,24,19,20,17,13,100,64,33,1021,1019,1016,10,7,3,13,8,,0.00,1,,113
2014-1-16,31,26,20,22,18,16,100,69,40,1022,1019,1017,10,7,2,16,8,,0.00,2,,114
2014-1-17,31,26,21,21,18,14,100,69,35,1021,1019,1016,8,6,3,14,11,,0.00,2,,111
2014-1-18,31,24,19,20,17,14,100,65,35,1021,1019,1016,8,5,1,14,10,,0.00,2,,112
```

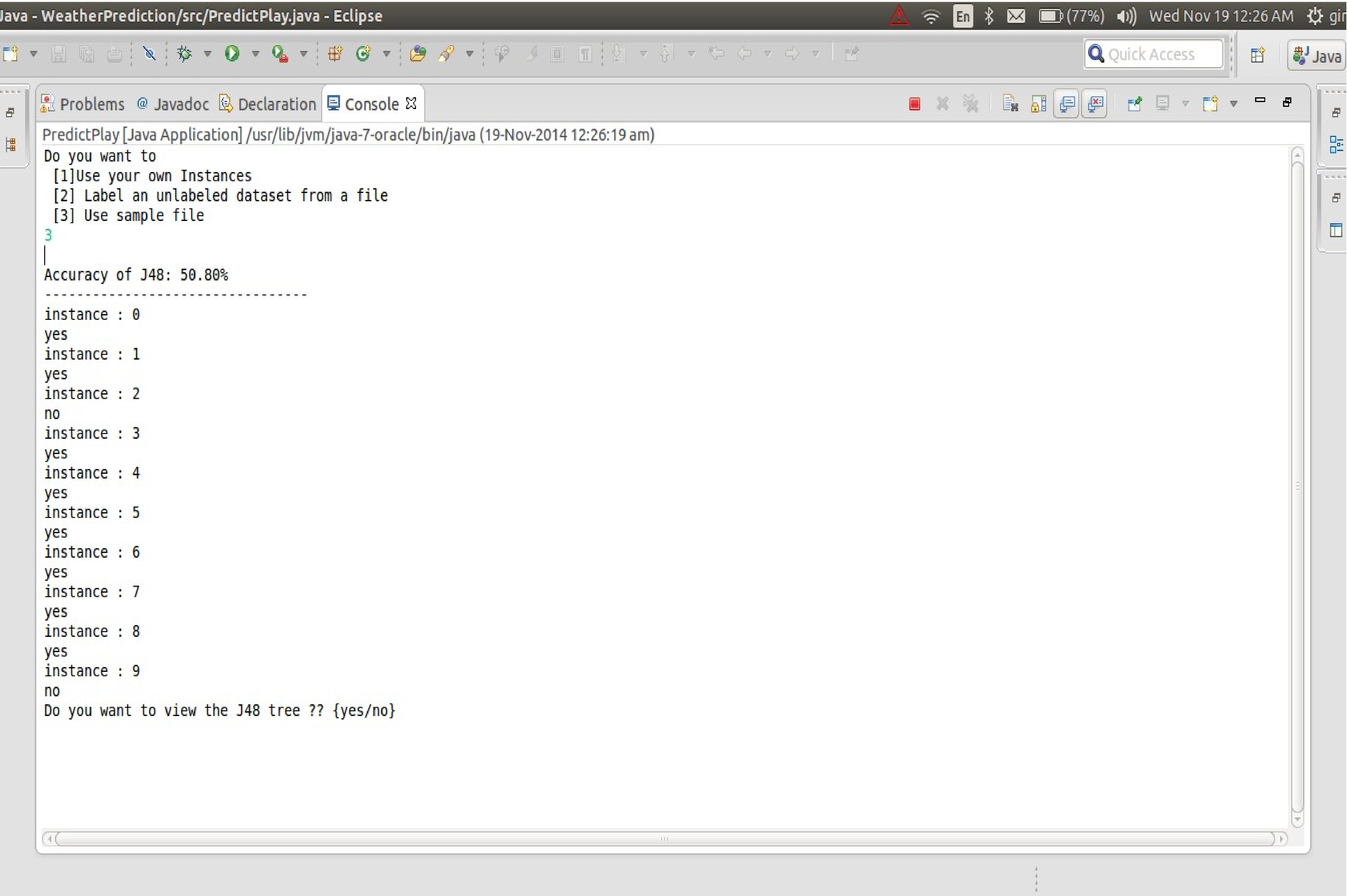
Sample Dataset
[arff format]



```
@relation weather

@attribute MaxTemp numeric
@attribute MeanTemp numeric
@attribute MinTemp numeric
@attribute MaxHumidity numeric
@attribute MeanHumidity numeric
@attribute MinHumidity numeric
@attribute WindSpeed numeric
@attribute CloudCover numeric
@attribute play {yes,no}
@data
29,24,19,100,76,45,23,3,no
29,24,20,100,74,45,14,4,yes
30,23,17,100,73,45,14,2,no
32,24,17,100,57,18,19,3,yes
32,24,17,100,62,24,16,2,yes
33,24,16,100,58,20,19,1,yes
31,24,18,100,61,25,21,3,no
30,24,18,100,67,37,21,3,no
31,24,18,100,73,38,23,3,yes
30,24,20,100,82,40,14,3,yes
32,27,22,73,44,27,13,3,yes
```

OUTPUT

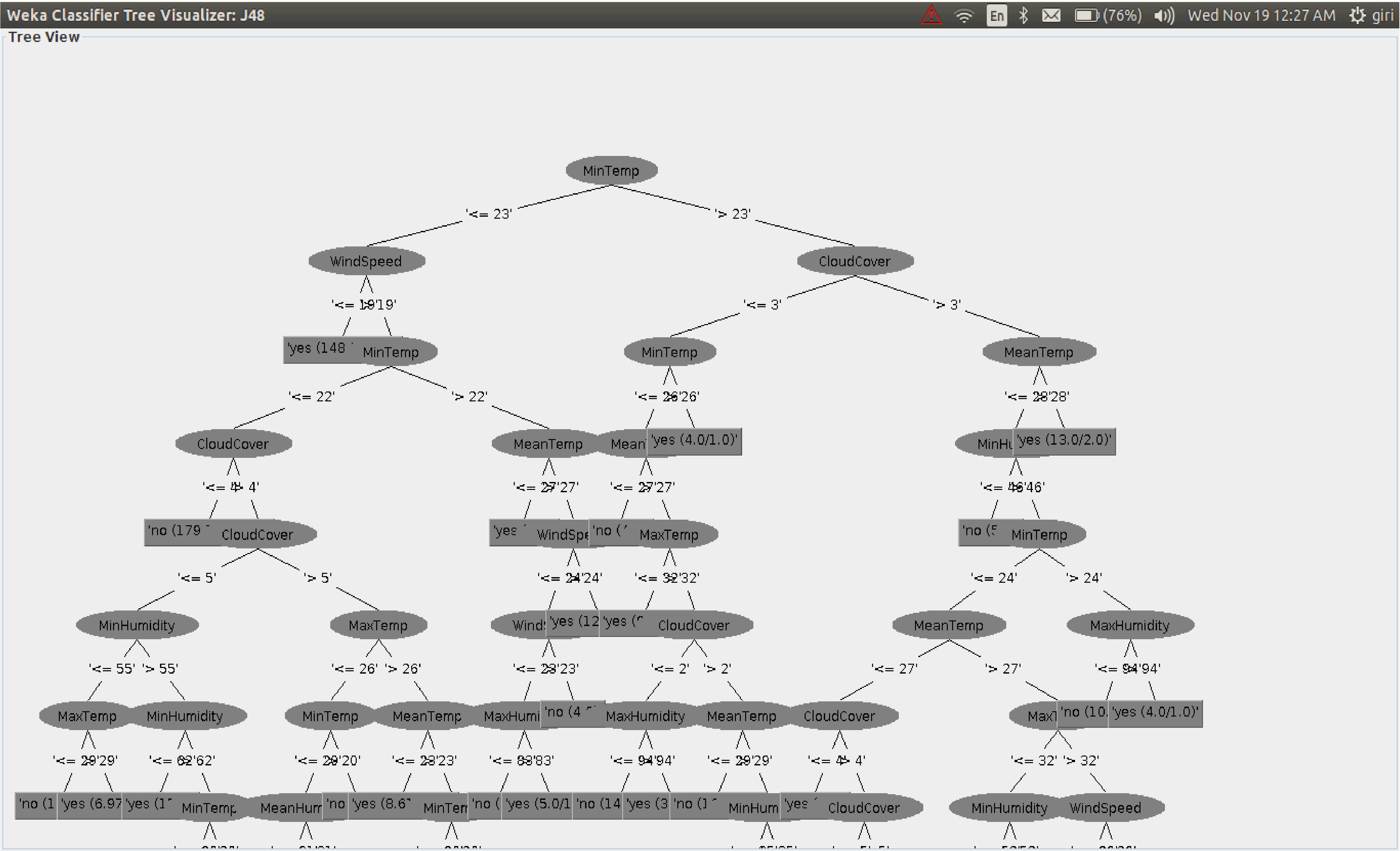


PredictPlay [Java Application] /usr/lib/jvm/java-7-oracle/bin/java (19-Nov-2014 12:26:19 am)

```
Do you want to
 [1]Use your own Instances
 [2] Label an unlabeled dataset from a file
 [3] Use sample file
3

Accuracy of J48: 50.80%
---------------------------------
instance : 0
yes
instance : 1
yes
instance : 2
no
instance : 3
yes
instance : 4
yes
instance : 5
yes
instance : 6
yes
instance : 7
yes
instance : 8
yes
instance : 9
no
Do you want to view the J48 tree ?? {yes/no}
```

OUTPUT


Weka Classifier Tree Visualizer: J48
(76%)
Wed Nov 19 12:27 AM
giri
Tree View
MinTemp
'<= 23'
'> 23'
WindSpeed
CloudCover
'<= 19'
'> 19'
'<= 3'
'> 3'
'yes (148
MinTemp
MinTemp
MeanTemp
'<= 22'
'> 22'
'<= 26'
'> 26'
'<= 28'
'> 28'
CloudCover
MeanTemp
Mean
'yes (4.0/1.0)'
MinHu
'yes (13.0/2.0)'
'<= 4'
'> 4'
'<= 27'
'> 27'
'<= 27'
'> 27'
'<= 46'
'> 46'
'no (179
CloudCover
'yes
WindSpe
'no (
MaxTemp
'no (5
MinTemp
'<= 5'
'> 5'
'<= 24'
'> 24'
'<= 32'
'> 32'
'<= 24'
'> 24'
MinHumidity
MaxTemp
Wind
'yes (12
'yes (
CloudCover
MeanTemp
MaxHumidity
'<= 55'
'> 55'
'<= 26'
'> 26'
'<= 23'
'> 23'
'<= 2'
'> 2'
'<= 27'
'> 27'
'<= 94'
'> 94'
MaxTemp
MinHumidity
MinTemp
MeanTemp
MaxHumi
'no (4
MaxHumidity
MeanTemp
CloudCover
MaxT
'no (10.
'yes (4.0/1.0)'
'<= 29'
'> 29'
'<= 62'
'> 62'
'<= 20'
'> 20'
'<= 23'
'> 23'
'<= 83'
'> 83'
'<= 94'
'> 94'
'<= 29'
'> 29'
'<= 4'
'> 4'
'<= 32'
'> 32'
'no (1
'yes (6.97
'yes (1
MinTemp
MeanHum
'no
'yes (8.6
MinTem
'no (
'yes (5.0/1
'no (14
'yes (3
'no (1
MinHum
'yes
CloudCover
MinHumidity
WindSpeed

RESULT

The decision tree was built using the training data set. The values for the 'play' attribute were predicted for each instance in the testing dataset according to the decision tree built.Thereafter, the accuracy of the prediction was calculated.

CONCLUSIONS

We concluded that using the concepts of machine learning one could easily predict values for a given attribute in a dataset if the values for the other attributes are known.

These prediction values could be further used for other purposes for e.g in our case the value of the play attribute could be used to decide whether a cricket match should be held in a particular city on a particular day or not given we have the weather data for that city.

References

[1] “Introduction to WEKA” *https://www.cs.auckland.ac.nz/courses/compsci367s1c/tutorials/IntroductionToWeka.pdf*

[2] Zdravko Markov,"An Introduction to the WEKA Data Mining System" , *http://www.cs.ccsu.edu/~markov/weka-tutorial.pdf*

[3]"Hello World program in WEKA " , *http://www.programcreek.com/2013/01/a-simple-machine-learning-example-in-java/*

[4] "Using WEKA in Java code" , *http://weka.wikispaces.com/Use+WEKA+in+your+Java+code*

[5] Ian H. Witten and Eibe Frank on "Introduction to Machine Learning", Data Mining :Practical Machine Learning

Thank You