2021

Data Warehousing And Data Mining

IT5003

Laboratory Journal

B.Tech (Information Technology)
5th Semester

Submitted to :
Department of Information Technology
C.G Patel Institute of Technology,
Bardoli

छोटुभाई गोपालभाई पटेल प्रौद्योगिकी संस्थान, बारडोली
Chhotubhai Gopalbhai Patel Institute of Technology, Bardoli

Certificate

This is to certify that

Mr/Ms. *Chitral Patel*

Enrolment No: 201903103510035 Exam Seat No. ________________

of

Semester 5th Course in

B.Tech(Information Technology)

has satisfactory completed his/her term work in

Data Warehousing And Data Mining (IT5003)

in the laboratory of this college during academic year 2021

Date of Submission: ________________

Prof. Monali Gandhi
Assistant Professor
Department of Information Technology,
C. G. Patel Institute of Technology, Uka Tarsadia University,
Bardoli – Surat

College Seal

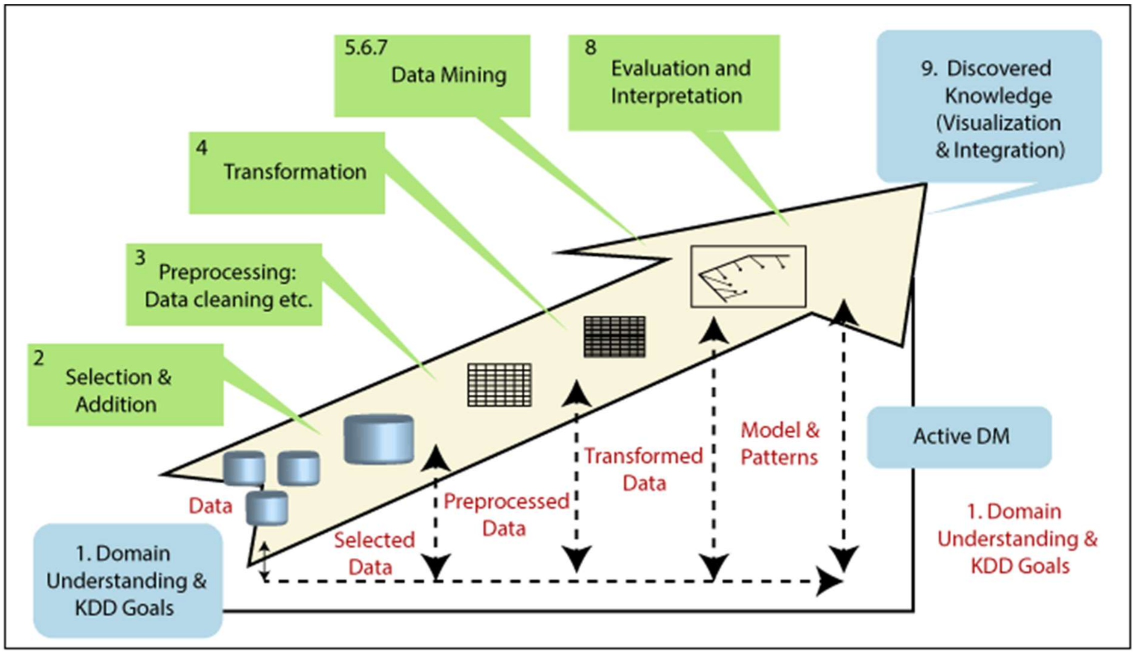
Prof. Purvi Tandel
Head of Department
Department of Information Technology,
C. G. Patel Institute of Technology, Uka Tarsadia University,
Bardoli – Surat.

PRACTICAL:1

AIM: Case study of any three data mining application and make a detailed note on them.

- **TELECOMMUNICATION INDUSTRY**:

CHART:


5.6.7
Data Mining
8
Evaluation and Interpretation
9. Discovered Knowledge (Visualization & Integration)
4
Transformation
3
Preprocessing: Data cleaning etc.
2
Selection & Addition
Data
Selected Data
Preprocessed Data
Transformed Data
Model & Patterns
Active DM
1. Domain Understanding & KDD Goals
1. Domain Understanding & KDD Goals

INTRODUCTION:

Expanding and growing at a fast pace, especially with the advent of the internet. Data mining can enable key industry players to improve their service quality to stay ahead in the game. Pattern analysis of spatiotemporal databases can play a huge role in mobile telecommunication, mobile computing, and also web and information services. And techniques like outlier analysis can detect fraudulent users. Also, OLAP and visualization tools can help compare information, such as user group behaviour, profit, data traffic, system overloads, etc.

BACKGROUND:

Telecommunication companies maintain an enormous amount of information about their customers and, due to an extremely competitive environment, have great motivation for exploiting this information. For these reasons the telecommunications industry has been a leader in the use of data mining to identify customers, retain customers, and maximize the profit obtained from each customer. Perhaps the most famous use of data mining to acquire new telecommunications customers was MCI's Friends and Family program. This program, long since retired, began after marketing researchers identified many small but wellconnected subgraphs in the graphs of calling activity (Han, Altman, Kumar, Mannila & Pregibon, 2002). By offering reduced rates to customers in one's calling circle, this marketing strategy enabled the company to use their own customers as salesmen. This work can be considered an early use of social-network analysis and link mining (Getoor & Diehl, 2005). A more recent example uses the interactions between consumers to identify those customers likely to adopt new telecommunication services (Hill, Provost & Volinsky, 2006). A more traditional approach involves generating customer profiles (i.e., signatures) from call detail records and then mining these profiles for marketing purposes. This approach has been used to identify whether a phone line is being used for voice or fax (Kaplan, Strauss & Szegedy, 1999) and to classify a phone line as belonging to a either business or residential customer (Cortes & Pregibon, 1998).

Main Focus:

Numerous data mining applications have been deployed in the telecommunications industry. However, most applications fall into one of the following three categories: marketing, fraud detection, and network fault isolation and prediction.

FUTURE TRENDS:

Data mining should play an important and increasing role in the telecommunications industry due to the large amounts of highquality data available, the competitive nature of the industry and the advances being made in data mining. In particular, advances in mining data streams, mining sequential and temporal data, and predicting/classifying rare events should benefit the telecommunications industry. As these and other advances are made, more reliance will be placed on the knowledge acquired through data mining and less on the knowledge acquired through the time-intensive process of eliciting domain knowledge from experts although we expect human experts will continue to play an important role for some time to come.

CONCLUSION:

The telecommunications industry has been one of the early adopters of data mining and has deployed numerous data mining applications. The primary applications relate to marketing, fraud detection, and network monitoring. Data mining in the telecommunications industry faces several challenges, due to the size of the data sets, the sequential and temporal nature of the data, and the real-time requirements of many of the applications. New methods have been developed and existing methods have been enhanced to respond to these challenges. The competitive and changing nature of the industry, combined with the fact that the industry generates enormous amounts of data, ensures that data mining will play an important role in the future of the telecommunications industry.

➢ FINANCIAL ANALYSIS:

INTRODUCTION:

The banking and finance industry relies on high-quality, reliable data. In loan markets, financial and user data can be used for a variety of purposes, like predicting loan payments and determining credit ratings. And data mining methods make such tasks more manageable. Classification techniques facilitate separation of crucial factors that influence customers' banking decisions from the irrelevant ones. Further, multidimensional clustering techniques allow identification of customers with similar loan payment behaviours. Data analysis and mining can also help detect money laundering and other financial crimes.

ABSTRACT:

Globalization has changed the phase of today's business world. As a result, to stay competitive in business entails the efficient use of modern tools to track past transaction records to analyze past business trend and future forecasts. Business Organizations faces different reactions and behavior from customers, which is partly due to insufficient information or the inability of a human analyst to understand the hidden pattern in business data. Customization of an investment portfolio for customers remain an enormous challenge to financial institutions even though there exist enough data in the financial market for analysis. Concern by the poor data usage by financial institutions in Cameroon, the purpose of this study was to address the problem of different distribution of business financial information and to provide solutions based on proper decision making from information support on studies of data warehousing technology of Bank of Cameroon for Credit and

Savings (BICEC). The comprehensive study included the overall planning of the system and the key technologies for achieving the system. Studies have showed that the financial decision-making system based on the data warehouse provides practical support for conducting the analysis of economic forecasts and policy research.

CONCLUSION:

After the financial crisis of 2008 and the global crisis of 2009, investors are becoming more cautious towards investments, especially in high risk financial products. These financial issues make it more difficult to devise a portfolio. In fact, customer orientation is becoming a trend in today's business. Many companies want to understand customers' needs, requirements and preferences in order to achieve high customer satisfaction. Customer satisfaction relies on superior products and services that the company provides. To customize the products and services, a company needs to gain more understanding of customer behavior. However, many companies lack a decision support system. The managers, thus, difficulty in understanding the correlations in the data and in customer behavior. In addition to providing superior products or services, many companies are facing the challenge of handling a huge growing amount of data in daily transactions. Attempting to address these challenges, the aim of the paper is to develop an intelligent Financial Data Mining Model (FDMM) that can help financial companies to tackle the problems. The fundamentals of FDMM are that firstly, all the relevant quality data are collected and preprocessed through DSPM, so each department can share information across the organization. Secondly, the CM is developed to partition the customers into specific groups. The segmented groups provide marketing implications to the sales managers so as to develop customer values for highly profitable customers.

➢HIGHER EDUCATION:

INTRODUCTION:

Nowadays, higher learning institutions encounter many problems, which keep them away from achieving their quality objectives. Most of these problems caused from knowledge gap. Knowledge gap is the lack of significant knowledge at the educational main processes such as advising, planning, registration, evaluation and marketing. For example, many learning institutions do not have access to the necessary information to advise students. Therefore, they are not able to give suitable recommendation for them. Data mining is a powerful technology that can be best defined as the automated process of extracting useful knowledge and information including, patterns, associations . The knowledge discovered by data mining techniques would enable the higher learning institutions in divers ways not limited to making better decisions, having more advanced planning in directing students, predicting individual behaviors with higher accuracy, and enabling the institution to allocate resources and staff more effectively. It results in improving the effectiveness and efficiency of the processes . One of the biggest challenges that higher education faces today is predicting the academic paths of student. Many higher education systems are unable detecting student population who are likely to drop out because of lack of intelligence method to use information and guidance from the university system.

CONCLUSION:

In this paper, we study the student dropout in computer science major in ALAQSA University for the purpose of improving the current teaching procedures and education strategies. Technologically, we do not propose new methods like FB-growth or decision-tree. However, we only use the mature classification and approaches in this study, cannot present more competitive algorithms or improve the existing algorithms. The study finds that mastering “digital design” and “algorithm analysis” courses has a great affect on predicting student persistence in the major and decrease student likelihood of dropout.

PRACTICAL:2

AIM: Write summarized description of any ten data mining tools.

1. Rapid Miner
2. Oracle Data Mining
3. IBM SPSS Modeler
4. Knime
5. Python
6. Orange
7. Kaggle
8. Rattle
9. Weka
10. Xplenty

➢ **Rapid Miner**

A data science software platform providing an integrated environment for various stages of data modelling including data preparation, data cleansing, exploratory data analysis, visualization and more. The techniques that the software helps with are machine learning, deep learning, text mining and predictive analytics. Easy to use GUI tools that take you through the modelling process. This tool written entirely in Java is an open-source framework and is wildly popular in the data mining world.

- ➢ **Oracle Data Mining:**

Oracle, the world leader it database software, combines it's prowess in database technologies with Analytical tools and brings you Oracle Advanced Analytics Database part of the Oracle

Enterprise Edition. It features several data mining algorithms for classification, regressing, prediction, anomaly detection and more. This is proprietary software and is supported by Oracle technical staff in helping your business build a robust data mining infrastructure at the enterprise scale.

- ➢ **IBM SPSS Modeler:**

IBM is again a big name in the data space when it comes to large enterprises. It combines well with leading technologies to implement a robust enterprise-wide solution. IBM SPSS Modeller is a visual data science and machine learning solution, helping in shortening the time to value by speeding up operational tasks for data scientists. IBM SPSS Modeler will have you covered from drag and drop data exploration to machine learning.

The software is used in leading enterprises for data preparation, discovery, predictive analytics, model management and deployment. The tool helps organizations to tap into their data assets and applications easily. One of the advantages of proprietary software is its ability to meet robust governance and security requirements of an organization at the enterprise level, and this reflects in every tool that IBM offers on the data mining front.

- ➢ **Knime**

Konstanz Information Miner is an open-source data analysis platform, helping you with build, deployment and scale in no time. The tool aims to help make predictive intelligence accessible to inexperience users. It aims to make the process easy by it is a step by step guide based GUI tools. The product markets itself as an End to End Data Science product, that helps create and production data science using its single easy and intuitive environment.

- ➢ **Python**

Python is a freely available and open-source language that is known to have a quick learning curve. Combined with is the ability as a general-purpose language and it is a large library of packages that help build a system for creating data models from the scratch, Python makes for a great tool for organizations who want the software they use to be custom built to their specifications.With Python, you won't get the fancy stuff that proprietary software offers, but the functionality is there for anybody to pick up and creates their own environment with graphical interfaces of their liking. What also supports python is the large online community of package developers who ensure the packages on offer are robust and secure. One of the features Python is known for in this field is powerful on the fly visualization features it offers.

➢ **Orange**

Orange is a machine learning and data science suite, using python scripting and visual programming featuring interactive data analysis and component-based assembly of data mining systems. Orange offers a broader range of features than most other Python-based data mining and machine learning tools. It is a software that has over 15 years of active development and use. Orange also offers a visual programming platform with GUI for interactive data visualization.

➢ **Kaggle**

The largest community of data scientists and machine learning professionals. Kaggle although started as a platform for machine learning competitions, is now extending its footprint into the public cloud-based data science platform arena. Kaggle now offers code and data that you need for your data science implementations. There are over 50k public datasets and 400k public notebooks that you can use to ramp up your data mining efforts. The huge online community that Kaggle enjoys is your safety net for implementation-specific challenges.

➢ **Rattle**

The rattle is an R language based GUI tool for data mining requirements. The tool is free and open-source and can be used to get statistical and visual summaries of data, the transformation of data for data models, build supervised and unsupervised machine learning models and compare model performance graphically.

➢ **Weka**

Waikato Environment for Knowledge Analysis (Weka) is a suite of machine learning tools written in Java. A collection of visualization tools for predictive modelling in a GUI presentation, helping you build your data models and test them, observing the model performances graphically.

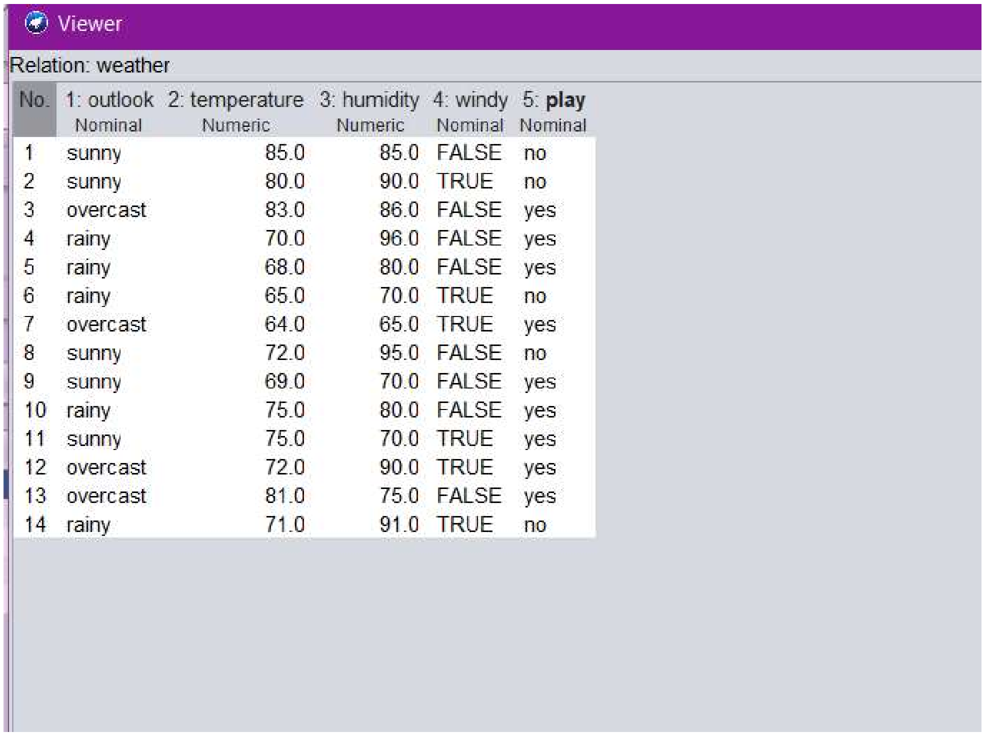
➢ **Xplenty**

Xplenty is a cloud-based data integration platform that helps read, process and prepare information from various databases and integrate it with a wide variety of business applications.

PRACTICAL:3

AIM: Perform the following tasks as per given instruction:
a. Apply Data Cleaning methods on given data set using Weka.
b. To perform data cleaning by handling missing values.
Pre processing for the missing value, by replacing them with all the following.
- With the global constant like Unknown or – infinity.
- Use the central tendency of attribute Mean (numerical attributes).
- Use the class wise attribute Mean or median.
- Apply all three normalization methods to any one numeric attribute.

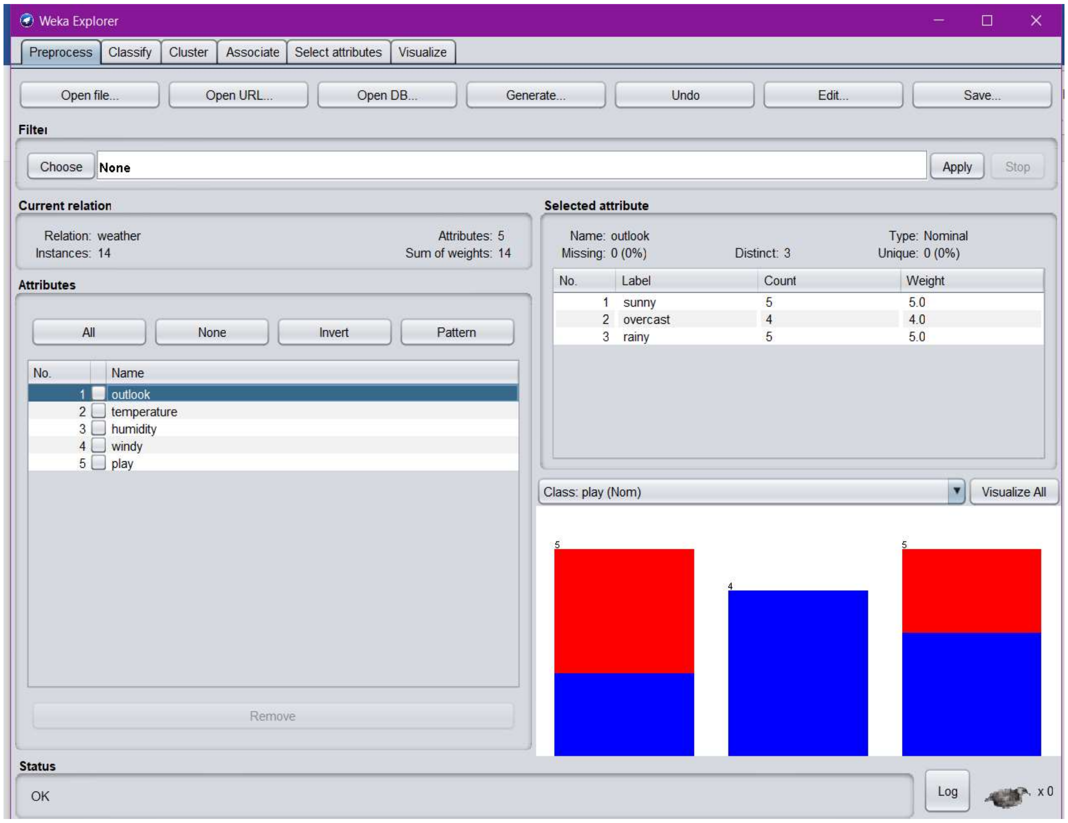
- Useless data are clean:



Relation: weather

No.	1: outlook Nominal	2: temperature Numeric	3: humidity Numeric	4: windy Nominal	5: play Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

- Graph represent clean of data:



Filter

None

Current relation

Relation: weather
Instances: 14
Attributes: 5
Sum of weights: 14

Attributes

No.	Name
1	outlook
2	temperature
3	humidity
4	windy
5	play

Selected attribute

Name: outlook
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)

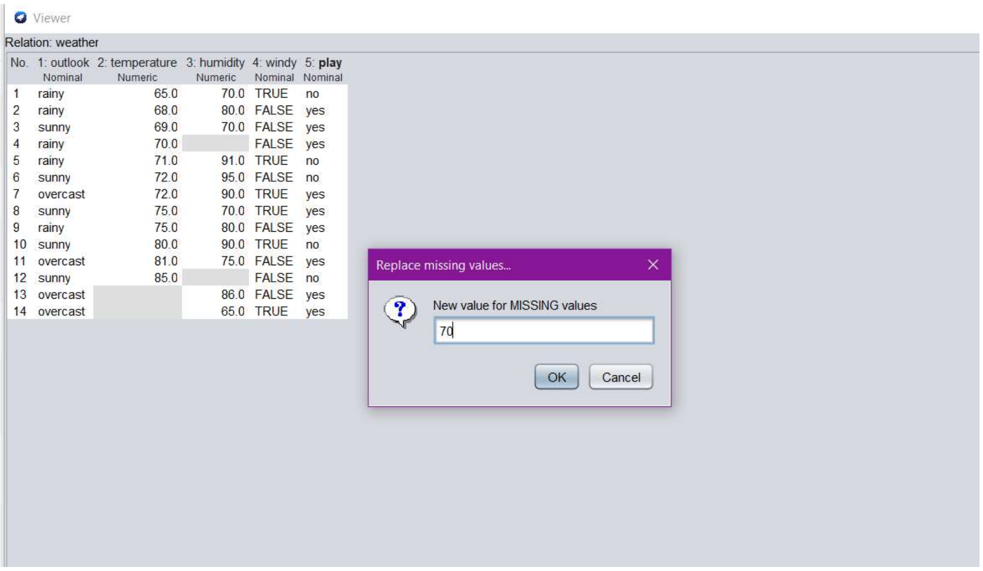
No.	Label	Count	Weight
1	sunny	5	5.0
2	overcast	4	4.0
3	rainy	5	5.0

Class: play (Nom)

Status

OK

- Missing value:



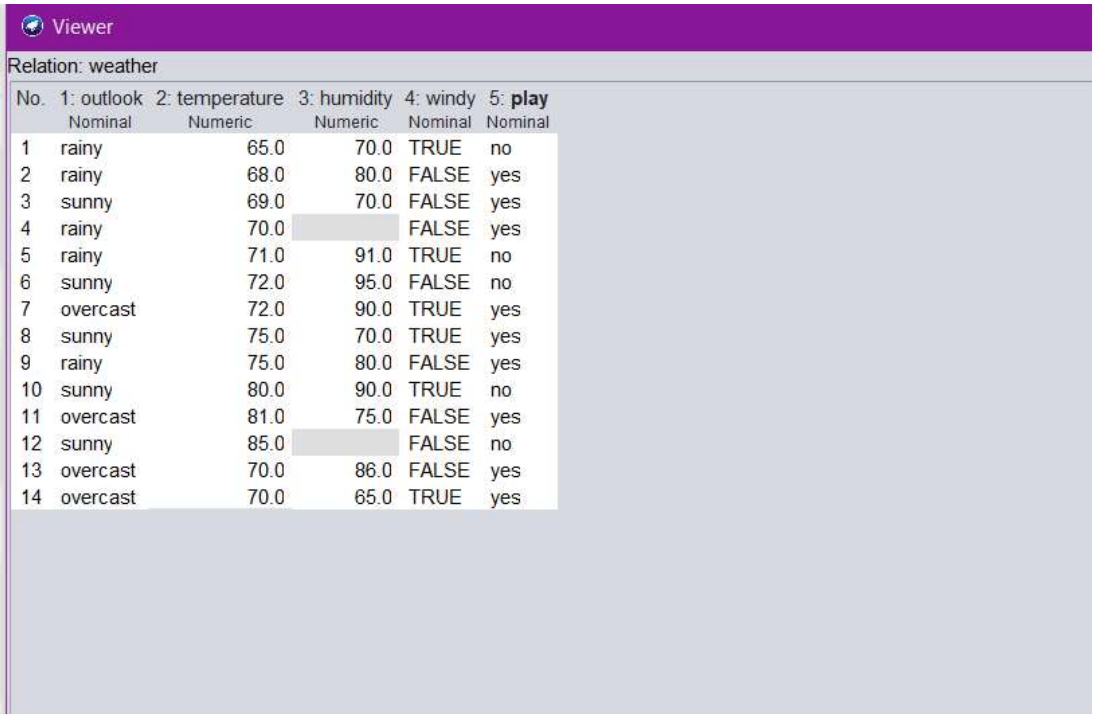
Relation: weather

No.	1: outlook Nominal	2: temperature Numeric	3: humidity Numeric	4: windy Nominal	5: **play** Nominal
1	rainy	65.0	70.0	TRUE	no
2	rainy	68.0	80.0	FALSE	yes
3	sunny	69.0	70.0	FALSE	yes
4	rainy	70.0		FALSE	yes
5	rainy	71.0	91.0	TRUE	no
6	sunny	72.0	95.0	FALSE	no
7	overcast	72.0	90.0	TRUE	yes
8	sunny	75.0	70.0	TRUE	yes
9	rainy	75.0	80.0	FALSE	yes
10	sunny	80.0	90.0	TRUE	no
11	overcast	81.0	75.0	FALSE	yes
12	sunny	85.0		FALSE	no
13	overcast		86.0	FALSE	yes
14	overcast		65.0	TRUE	yes

Replace missing values...

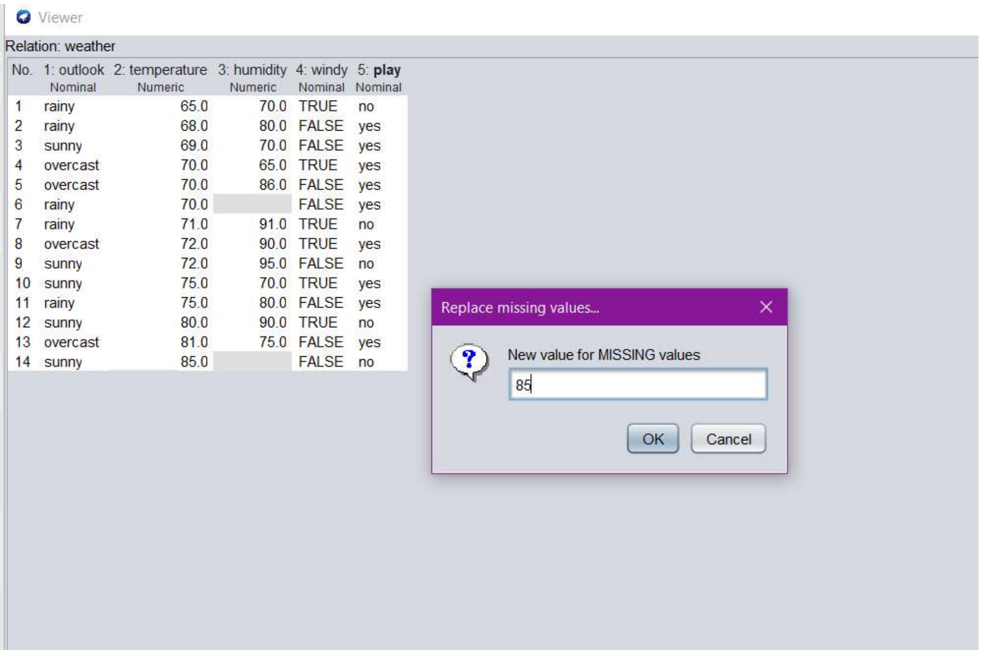
New value for MISSING values

70



Relation: weather

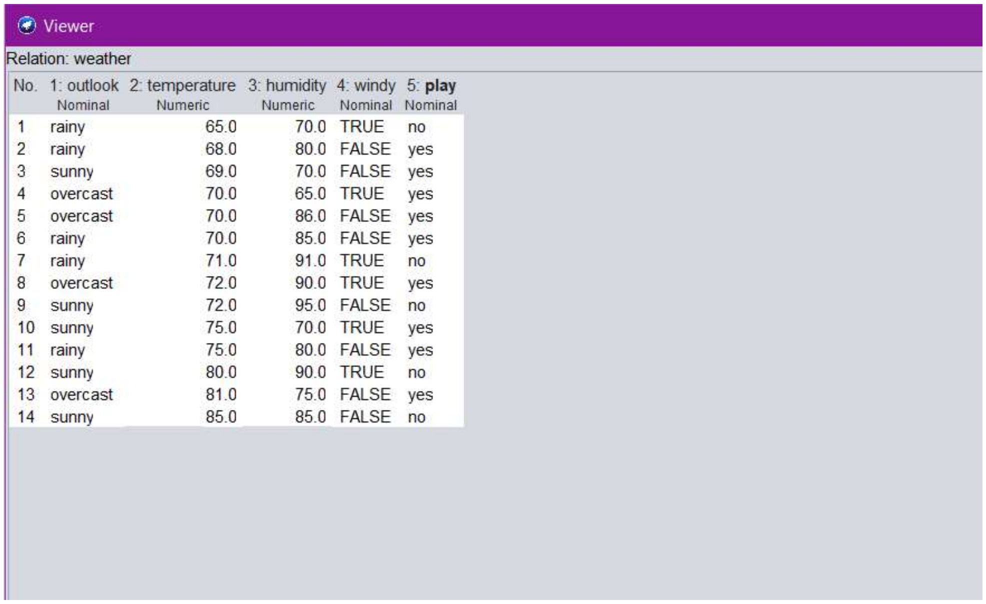
No.	1: outlook Nominal	2: temperature Numeric	3: humidity Numeric	4: windy Nominal	5: **play** Nominal
1	rainy	65.0	70.0	TRUE	no
2	rainy	68.0	80.0	FALSE	yes
3	sunny	69.0	70.0	FALSE	yes
4	rainy	70.0		FALSE	yes
5	rainy	71.0	91.0	TRUE	no
6	sunny	72.0	95.0	FALSE	no
7	overcast	72.0	90.0	TRUE	yes
8	sunny	75.0	70.0	TRUE	yes
9	rainy	75.0	80.0	FALSE	yes
10	sunny	80.0	90.0	TRUE	no
11	overcast	81.0	75.0	FALSE	yes
12	sunny	85.0		FALSE	no
13	overcast	70.0	86.0	FALSE	yes
14	overcast	70.0	65.0	TRUE	yes



Relation: weather

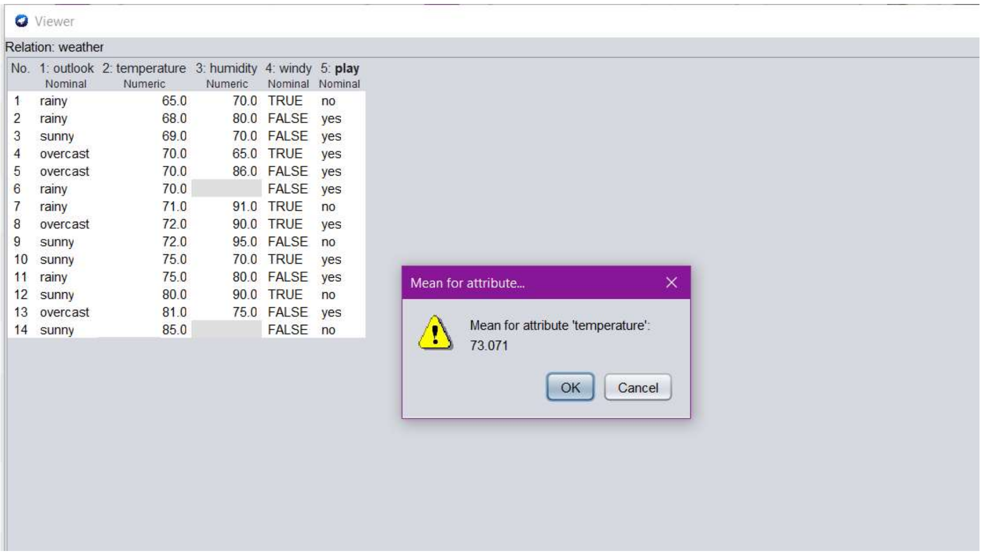
No.	1: outlook Nominal	2: temperature Numeric	3: humidity Numeric	4: windy Nominal	5: **play** Nominal
1	rainy	65.0	70.0	TRUE	no
2	rainy	68.0	80.0	FALSE	yes
3	sunny	69.0	70.0	FALSE	yes
4	overcast	70.0	65.0	TRUE	yes
5	overcast	70.0	86.0	FALSE	yes
6	rainy	70.0		FALSE	yes
7	rainy	71.0	91.0	TRUE	no
8	overcast	72.0	90.0	TRUE	yes
9	sunny	72.0	95.0	FALSE	no
10	sunny	75.0	70.0	TRUE	yes
11	rainy	75.0	80.0	FALSE	yes
12	sunny	80.0	90.0	TRUE	no
13	overcast	81.0	75.0	FALSE	yes
14	sunny	85.0		FALSE	no

New value for MISSING values: 85



Relation: weather

No.	1: outlook Nominal	2: temperature Numeric	3: humidity Numeric	4: windy Nominal	5: **play** Nominal
1	rainy	65.0	70.0	TRUE	no
2	rainy	68.0	80.0	FALSE	yes
3	sunny	69.0	70.0	FALSE	yes
4	overcast	70.0	65.0	TRUE	yes
5	overcast	70.0	86.0	FALSE	yes
6	rainy	70.0	85.0	FALSE	yes
7	rainy	71.0	91.0	TRUE	no
8	overcast	72.0	90.0	TRUE	yes
9	sunny	72.0	95.0	FALSE	no
10	sunny	75.0	70.0	TRUE	yes
11	rainy	75.0	80.0	FALSE	yes
12	sunny	80.0	90.0	TRUE	no
13	overcast	81.0	75.0	FALSE	yes
14	sunny	85.0	85.0	FALSE	no

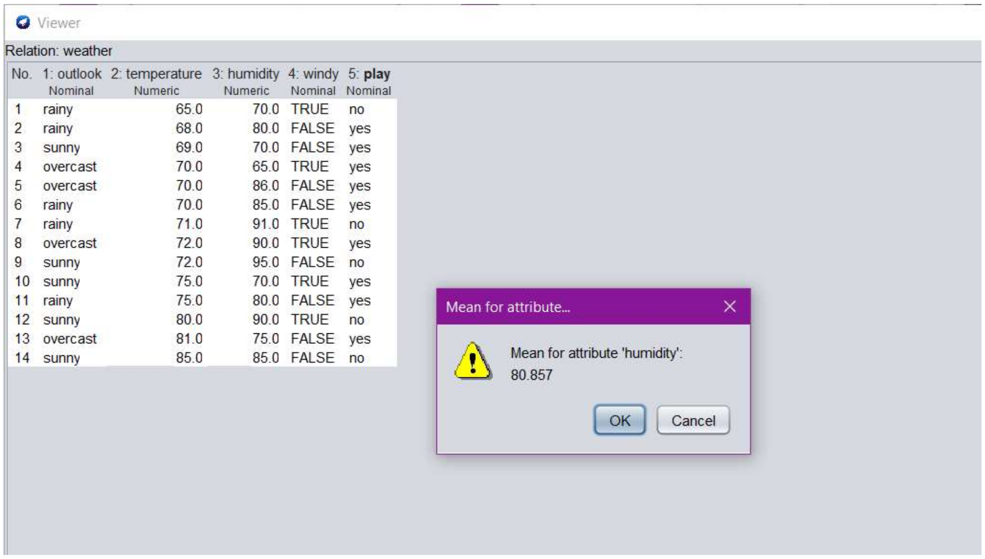


Relation: weather

No.	1: outlook	2: temperature	3: humidity	4: windy	5: **play**
	Nominal	Numeric	Numeric	Nominal	Nominal
1	rainy	65.0	70.0	TRUE	no
2	rainy	68.0	80.0	FALSE	yes
3	sunny	69.0	70.0	FALSE	yes
4	overcast	70.0	65.0	TRUE	yes
5	overcast	70.0	86.0	FALSE	yes
6	rainy	70.0		FALSE	yes
7	rainy	71.0	91.0	TRUE	no
8	overcast	72.0	90.0	TRUE	yes
9	sunny	72.0	95.0	FALSE	no
10	sunny	75.0	70.0	TRUE	yes
11	rainy	75.0	80.0	FALSE	yes
12	sunny	80.0	90.0	TRUE	no
13	overcast	81.0	75.0	FALSE	yes
14	sunny	85.0		FALSE	no

Mean for attribute...

Mean for attribute 'temperature':
73.071



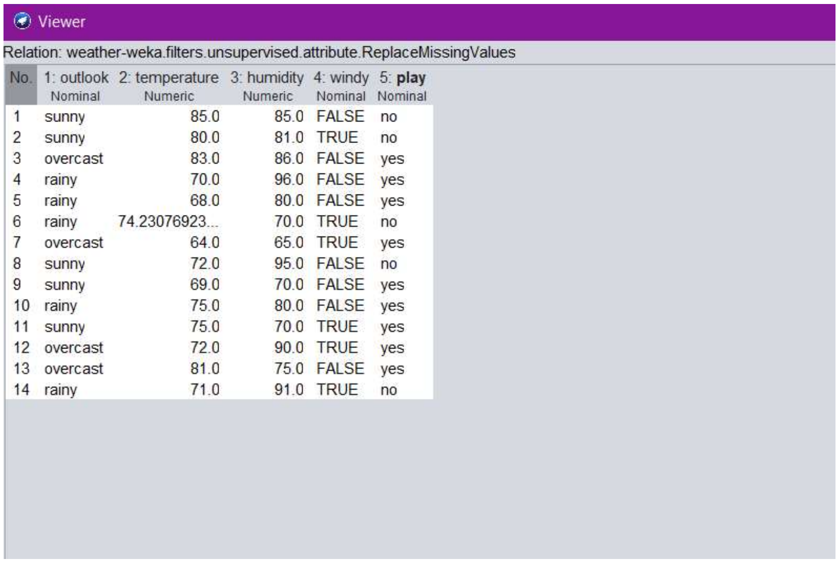
Relation: weather

No.	1: outlook	2: temperature	3: humidity	4: windy	5: **play**
	Nominal	Numeric	Numeric	Nominal	Nominal
1	rainy	65.0	70.0	TRUE	no
2	rainy	68.0	80.0	FALSE	yes
3	sunny	69.0	70.0	FALSE	yes
4	overcast	70.0	65.0	TRUE	yes
5	overcast	70.0	86.0	FALSE	yes
6	rainy	70.0	85.0	FALSE	yes
7	rainy	71.0	91.0	TRUE	no
8	overcast	72.0	90.0	TRUE	yes
9	sunny	72.0	95.0	FALSE	no
10	sunny	75.0	70.0	TRUE	yes
11	rainy	75.0	80.0	FALSE	yes
12	sunny	80.0	90.0	TRUE	no
13	overcast	81.0	75.0	FALSE	yes
14	sunny	85.0	85.0	FALSE	no

Mean for attribute...

Mean for attribute 'humidity':
80.857

- Null value is replace by written Null or infinte value:

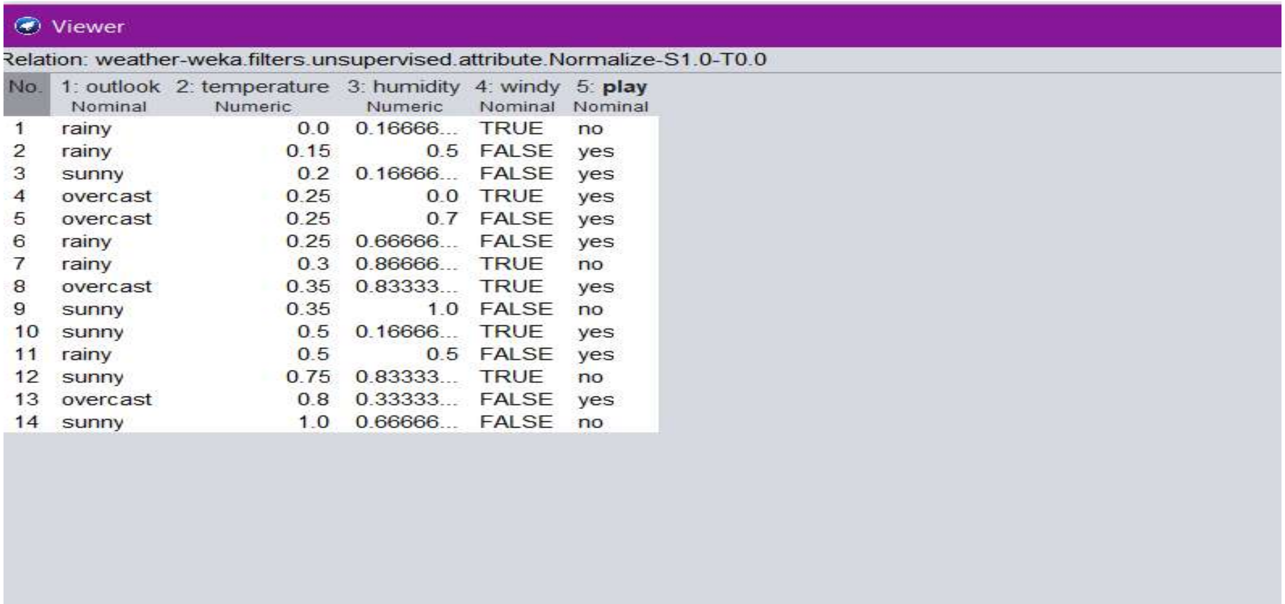


Viewer

Relation: weather-weka.filters.unsupervised.attribute.ReplaceMissingValues

No.	1: outlook Nominal	2: temperature Numeric	3: humidity Numeric	4: windy Nominal	5: **play** Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	81.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	74.23076923...	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

- Normalize data:

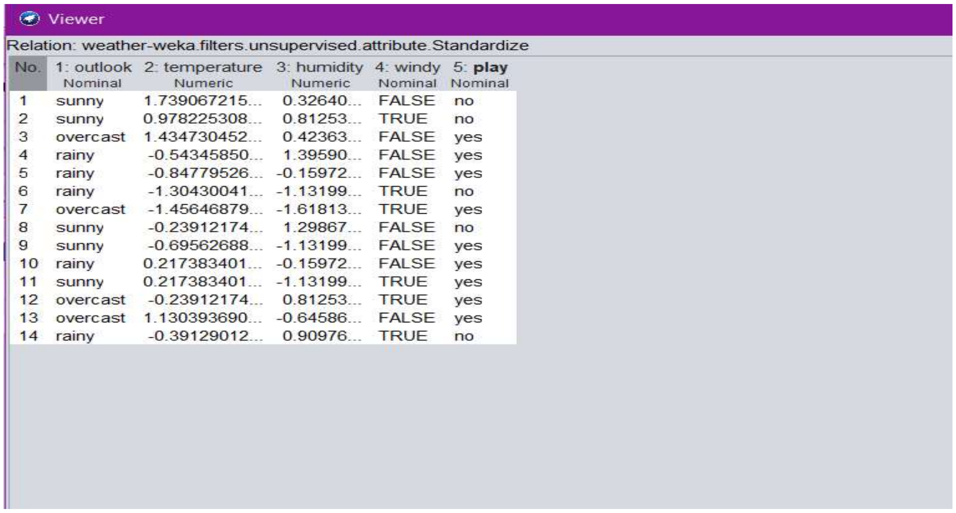


Viewer

Relation: weather-weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0

No.	1: outlook Nominal	2: temperature Numeric	3: humidity Numeric	4: windy Nominal	5: **play** Nominal
1	rainy	0.0	0.16666...	TRUE	no
2	rainy	0.15	0.5	FALSE	yes
3	sunny	0.2	0.16666...	FALSE	yes
4	overcast	0.25	0.0	TRUE	yes
5	overcast	0.25	0.7	FALSE	yes
6	rainy	0.25	0.66666...	FALSE	yes
7	rainy	0.3	0.86666...	TRUE	no
8	overcast	0.35	0.83333...	TRUE	yes
9	sunny	0.35	1.0	FALSE	no
10	sunny	0.5	0.16666...	TRUE	yes
11	rainy	0.5	0.5	FALSE	yes
12	sunny	0.75	0.83333...	TRUE	no
13	overcast	0.8	0.33333...	FALSE	yes
14	sunny	1.0	0.66666...	FALSE	no

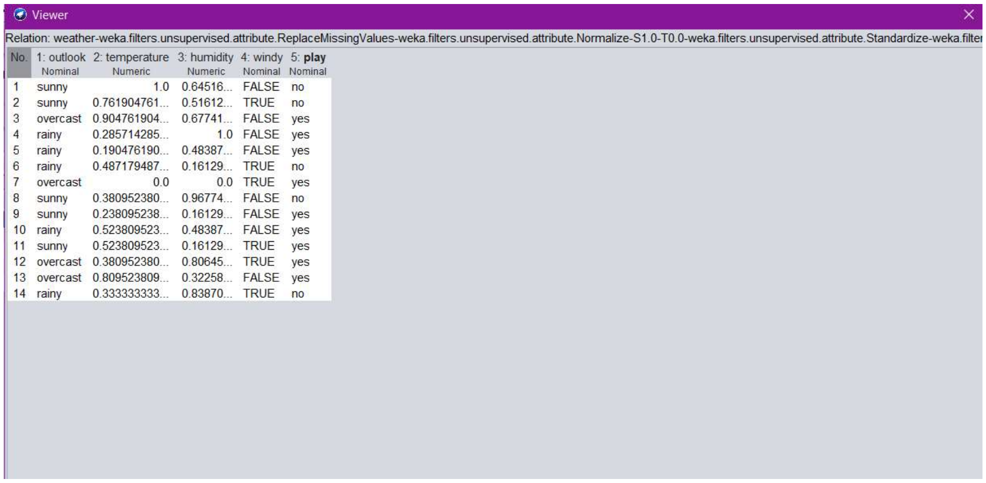
- Standardization:



Relation: weather-weka.filters.unsupervised.attribute.Standardize

No.	1: outlook Nominal	2: temperature Numeric	3: humidity Numeric	4: windy Nominal	5: **play** Nominal
1	sunny	1.739067215...	0.32640...	FALSE	no
2	sunny	0.978225308...	0.81253...	TRUE	no
3	overcast	1.434730452...	0.42363...	FALSE	yes
4	rainy	-0.54345850...	1.39590...	FALSE	yes
5	rainy	-0.84779526...	-0.15972...	FALSE	yes
6	rainy	-1.30430041...	-1.13199...	TRUE	no
7	overcast	-1.45646879...	-1.61813...	TRUE	yes
8	sunny	-0.23912174...	1.29867...	FALSE	no
9	sunny	-0.69562688...	-1.13199...	FALSE	yes
10	rainy	0.217383401...	-0.15972...	FALSE	yes
11	sunny	0.217383401...	-1.13199...	TRUE	yes
12	overcast	-0.23912174...	0.81253...	TRUE	yes
13	overcast	1.130393690...	-0.64586...	FALSE	yes
14	rainy	-0.39129012...	0.90976...	TRUE	no

- Mathe Expression:



Relation: weather-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0-weka.filters.unsupervised.attribute.Standardize-weka.filter

No.	1: outlook Nominal	2: temperature Numeric	3: humidity Numeric	4: windy Nominal	5: **play** Nominal
1	sunny	1.0	0.64516...	FALSE	no
2	sunny	0.761904761...	0.51612...	TRUE	no
3	overcast	0.904761904...	0.67741...	FALSE	yes
4	rainy	0.285714285...	1.0	FALSE	yes
5	rainy	0.190476190...	0.48387...	FALSE	yes
6	rainy	0.487179487...	0.16129...	TRUE	no
7	overcast	0.0	0.0	TRUE	yes
8	sunny	0.380952380...	0.96774...	FALSE	no
9	sunny	0.238095238...	0.16129...	FALSE	yes
10	rainy	0.523809523...	0.48387...	FALSE	yes
11	sunny	0.523809523...	0.16129...	TRUE	yes
12	overcast	0.380952380...	0.80645...	TRUE	yes
13	overcast	0.809523809...	0.32258...	FALSE	yes
14	rainy	0.333333333...	0.83870...	TRUE	no