S. B. JAIN INSTITUTE OF TECHNOLOGY, MANAGEMENT
& RESEARCH, NAGPUR.
(An Autonomous Institute, Affiliated to RTMNU, Nagpur)
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
"To become a center for quality education in the field of computer science & engineering and to create competent professionals."

Session 2022-2023

Data Warehousing and Mining Lab (BECSE401P)

LAB MANUAL

Year: 4th Year

Semester: 7th Semester

Course Outcomes (COs): Upon the completion of the lab students will be able to

Subject	Course Outcomes	
Data Warehousing & Mining (Practical)	**CO401P.1**	Analyze the concept of Data Mining with Weka tool and its attributes
	CO401P.2	Build .arff and make use of Weka Tool
	CO401P.3	Analyze the various Mining Techniques in Weka as well as solve basic statistical calculations on data
	CO401P.4	Evaluate the aspect of data preprocessing, data cleaning and integration on various data set.
	CO401P.5	Apply and analyze Decision Trees and Clustering Rules (Supervised Approach).

Sr. No.	Aim of Practical	Unit No	CO Mapped
1	Introduction to Weka Tool & Creation of .arff file	1,2	**CO1, CO2**
2	Discretize the attribute using WEKA TOOL.	1,2	**CO1, CO2**
3	Demonstration of pre-processing on .arff as well as .CSV.	1	**CO3, CO4**
4	Demonstration of Association Rule process on dataset using Apriori Algorithm.	4	**CO3**
5	Demonstration of classification rule process on dataset using Naive Baye's Algorithm.	3	**CO3**
6	Demonstration of classification rule process on dataset using BayesNet Algorithm.	4	**CO3**
7	Demonstration of Classification Rule process on dataset using j48 algorithm.	4	**CO3**
8	Demonstration of classification rule process on dataset using random tree algorithm.	4	**CO3**
9	Demonstration of clustering rule process on dataset using DBSCAN.	5	**CO5**
10	Demonstration of clustering rule process on dataset using simple k-means	5	**CO5**
11	Introduction to new mining tool rapidminer.	5	**CO1, CO2**

Practical No.1

Aim: Introduction to Weka tool & Creation of .arff file

AIM: Introduction to Weka tool & Creation of .arff file.

OBJECTIVES:

- To learn how to explore WEKA tool.
- To create attribute file and distinguish database with data set.
- To analyze the data set for preprocessing.

AIM: Introduction to Weka tool & Creation of .arff file.

OBJECTIVES:

- To learn how to explore WEKA tool.
- To create attribute file and distinguish database with data set.
- To analyze the data set for preprocessing.

THEORY: The Weka GUI Chooser (class weka.gui.GUIChooser) provides a starting point for launching Weka's main GUI applications and supporting tools. If one prefers a MDI ("multiple document interface") appearance, then this is provided by an alternative launcher called "Main" (class weka.gui.Main).The GUI Chooser consists of four buttons—one for each of the four major Weka applications—and four menus.The buttons can be used to start the following applications:

• Explorer An environment for exploring data with WEKA (the rest of this documentation deals with this application in more detail).

• Experimenter An environment for performing experiments and conducting statistical tests between learning schemes.

• Knowledge Flow This environment supports essentially the same functions as the Explorer but with a drag-and-drop interface. One advantage is that it supports incremental learning.

• Workbench An all-in-one application that combines all the others within user-selectable "perspectives".

• Simple CLI Provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface.

The menu consists of four sections:

1. Program
2. Tools Other useful applications.
3. Visualization Ways of visualizing data with WEKA.

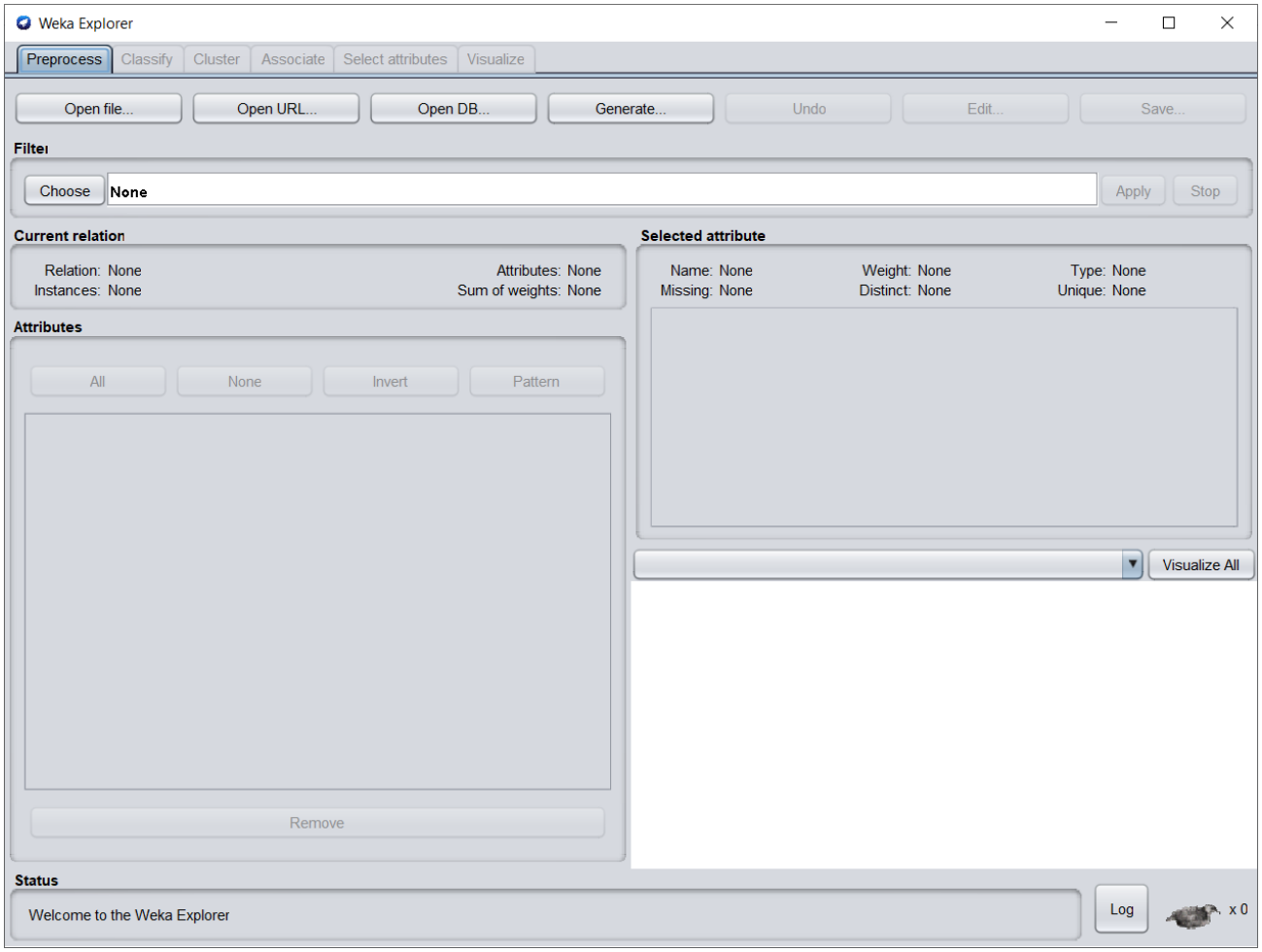
4. Help Online resources for WEKA can be found here.

To make it easy for the user to add new functionality to the menu without having to modify the code of WEKA itself, the GUI now offers a plugin mechanism for such add-ons. Due to the inherent dynamic class discovery, plugins only need to implement the weka.gui.MainMenuExtension interface and WEKA notified of the package they reside in to be displayed in the menu under “Extensions” (this extra menu appears automatically as soon as extensions are discovered).

If you launch WEKA from a terminal window, some text begins scrolling in the terminal. Ignore this text unless something goes wrong, in which case it can help in tracking down the cause

CODE: If Applicable

OUTPUT:



CONCLUSION: Thus, we successfully created .arff file using a weka tool.

DISCUSSION AND VIVA VOCE:

- What is expansion of .arff?

  Files with **arff extension** are related to **Attribute-Relation File Format**.

- How dataset is different than database?

  A **dataset** is a structured collection of data generally associated with a unique body of work while a **database** is an organized collection of data stored as *multiple* datasets. Those datasets are generally stored and accessed electronically from a computer system that allows the data to be easily accessed, manipulated, and updated.

- What is data?

  Data is a collection of facts, such as numbers, words, measurements, observations or just descriptions of things.

- What is information?

  Information is a sequence of symbols that carries a message, a set of items in which meaning is conveyed, or a specified arrangement of complex structures that convey a message to a

receiver.

- Explain various steps to launch WEKA?

To install WEKA on your machine, visit WEKA's official website and download the installation file. WEKA supports installation on Windows, Mac OS X and Linux. You just need to follow the instructions on this page to install WEKA for your OS.

The steps for installing on Mac are as follows:

- Download the Mac installation file.
- Double click on the downloaded weka-3-8-3-corretto-jvm.dmg file.
- You will see the following screen on successful installation.
- Click on the weak-3-8-3-corretto-jvm icon to start Weka.
- Optionally you may start it from the command line:
- java -jar weka.jar
- The WEKA GUI Chooser application will start and you would see the following screen:

The GUI Chooser application allows you to run five different types of applications as listed here:

- Explorer
- Experimenter
- KnowledgeFlow
- Workbench
- Simple CLI

REFERENCE:

http://www.srmuniv.ac.in

www.cs.sfu.ca/~han/DMbook.html

http://www.cs.waikato.ac.nz/ml/weka/documentation.html

https://publish.pothi.com/preview/?sku=ebook3591

http://www.cs.ccsu.edu/~markov/weka-tutorial.pdf

Data Mining – Concepts and Techniques, Jiawei Han & Micheline Kamber, Morgan Kauf-mann Publishers, Elsevier, 2nd Edition, 2006.

Practical No.2

Aim: Discretize the attribute using WEKA TOOL

AIM: Discretize the attribute using WEKA TOOL.

OBJECTIVES:

The objectives and expected learning outcomes of this practical are:

- To explain the Discretizing concept.
- Partitioned the selected attribute into given number of bins.

AIM: Discretize the attribute using WEKA TOOL.

OBJECTIVES:

- To explain the Discretizing concept.
- Partitioned the selected attribute into given number of bins.

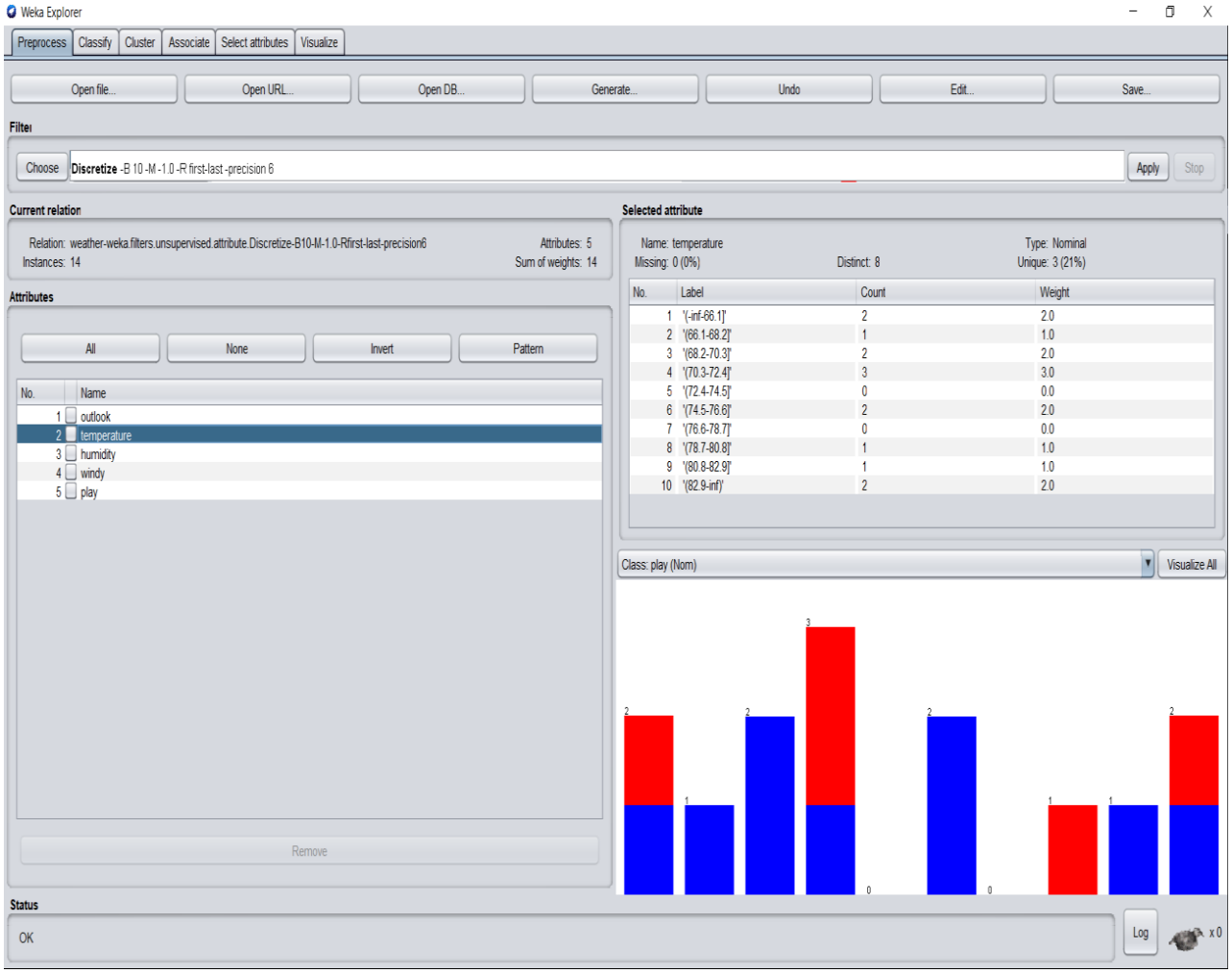
THEORY:

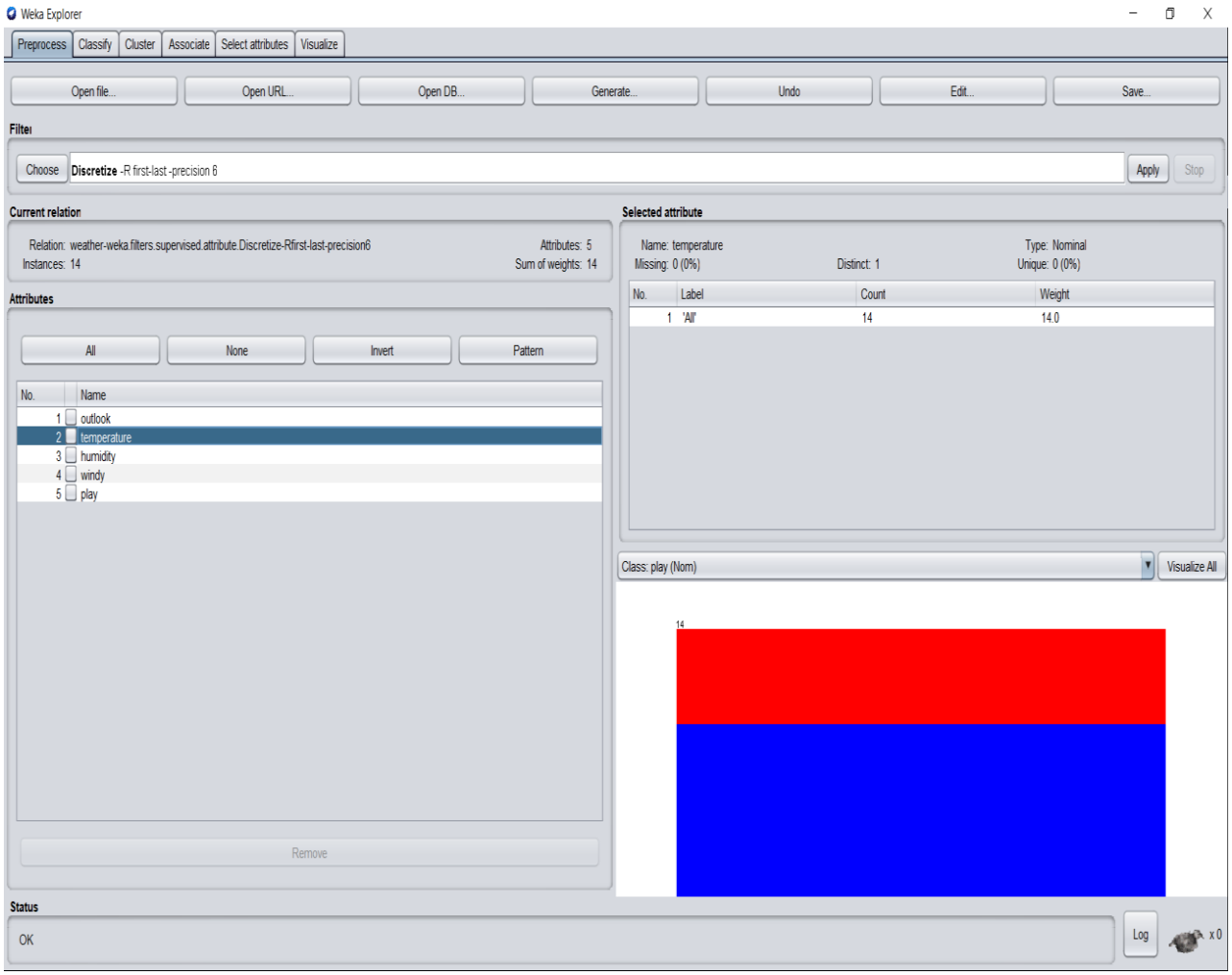
Discretization is the process of tranformation numeric data into nominal data, by putting the numeric values into distinct groups.

Common Approaches:

Unsupervised
Equal-Width Binning:
It divides the scope of possible values into N subscopes (bins) of the samewidth
Equal-Frequency Binning:
It divides the scope of possible values into N subscopes where each subscope (bin) carries the same number of instances
Supervised – classes are taken into account

OUTPUT:


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
Discretize -B 10 -M -1.0 -R first-last -precision 6
Apply
Stop
Current relation
Relation: weather-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-precision6
Instances: 14
Attributes: 5
Sum of weights: 14
Attributes
All
None
Invert
Pattern
No.
Name
1 outlook
2 temperature
3 humidity
4 windy
5 play
Remove
Selected attribute
Name: temperature
Missing: 0 (0%)
Distinct: 8
Type: Nominal
Unique: 3 (21%)
No.
Label
Count
Weight
1 '(-inf-66.1]' 2 2.0
2 '(66.1-68.2]' 1 1.0
3 '(68.2-70.3]' 2 2.0
4 '(70.3-72.4]' 3 3.0
5 '(72.4-74.5]' 0 0.0
6 '(74.5-76.6]' 2 2.0
7 '(76.6-78.7]' 0 0.0
8 '(78.7-80.8]' 1 1.0
9 '(80.8-82.9]' 1 1.0
10 '(82.9-inf)' 2 2.0
Class: play (Nom)
Visualize All
Status
OK
Log
x 0



CONCLUSION: Thus, we learnt to discretize the attribute using WEKA TOOL.

DISCUSSION AND VIVA VOCE:

- What is preprocessing?

  Data preprocessing is a step in the data mining and data analysis process that takes raw data and transforms it into a format that can be understood and analyzed by computers and machine learning.

- Why we need to preprocess the data?

  When using data sets to train machine learning models, you'll often hear the phrase **"garbage in, garbage out"** This means that if you use bad or "dirty" data to train your model, you'll end up with a bad, improperly trained model that won't actually be relevant to your analysis.

- How to perform data cleaning?

Data cleaning is the process of adding missing data and correcting, repairing, or removing incorrect or irrelevant data from a data set. Dating cleaning is the most important step of pre-processing because it will ensure that your data is ready to go for your downstream needs.

Data cleaning will correct all of the inconsistent data you uncovered in your data quality assessment. Depending on the kind of data you're working with, there are a number of possible cleaners you'll need to run your data through.

- Expand .csv?

  Comma-Separated Values (CSV) is a file format used to store tabular data in which numbers and text are stored in a plain-text form that can be easily written and read in a text editor. CSV format is the most common import and export format for spreadsheets and databases.

- Differentiate .arff and .csv?

  ARFF is an acronym that stands for Attribute-Relation File Format. It is an extension of the CSV file format where a header is used that provides metadata about the data types in the columns.

REFERENCE:

- *http://ai.fon.bg.ac.rs/wp-content/uploads/2015/04/ML-Attribute-Discretisation-and-Selection-Clustering-2014_eng.pdf*
- *https://weka.wikispaces.com/Discretizing+datasets*
- *Data Mining – Concepts and Techniques, Jiawei Han & Micheline Kamber, Morgan Kaufmann Publishers, Elsevier, 2nd Edition, 2006.*
- *https://publish.pothi.com/preview/?sku=ebook3591*
- *http://www.cs.ccsu.edu/~markov/weka-tutorial.pdf*

Practical No.3

Aim: Demonstration of pre-processing on .arff as well as .CSV.

AIM: Perform Preprocessing (data cleansing) on .arff as well as .csv format

OBJECTIVES:

- To explain the preprocessing concept.
- To apply data cleansing methods (removal of attributes, formation of bins) on .arff & .csv file.

AIM: Perform Preprocessing(data discretization) on .arff as well as .csv format.

OBJECTIVES:

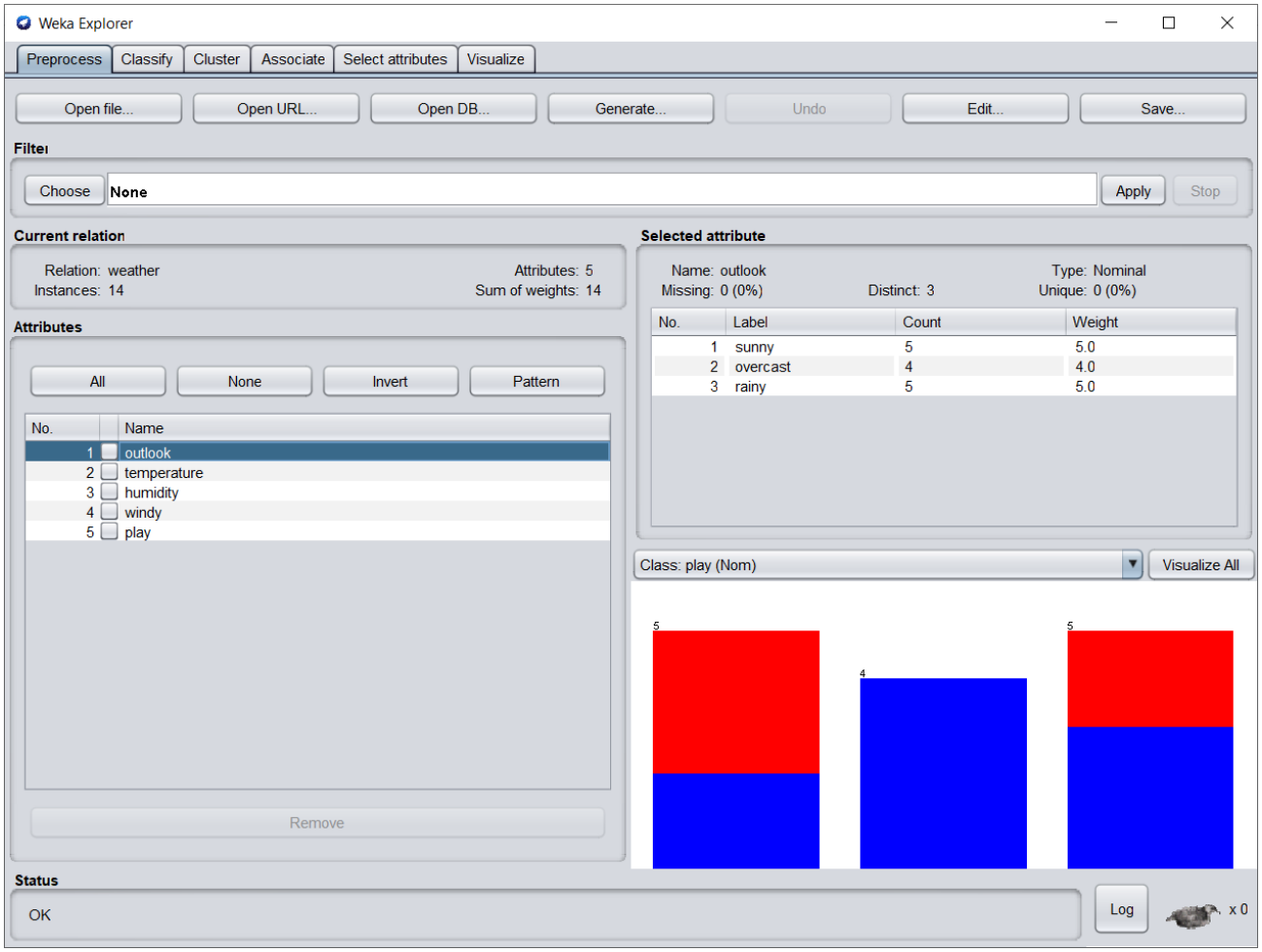
- To explain the preprocessing concept.
- To apply data cleansing methods(discretization of attributes, formation of bins) on .arff & .csv file.

THEORY:

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing Data goes through a series of steps during preprocessing:

- **Data Cleaning:** Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.
- **Data Integration:** Data with different representations are put together and conflicts within the data are resolved.
- **Data Transformation:** Data is normalized, aggregated and generalized.
- **Data Reduction:** This step aims to present a reduced representation of the data in a data warehouse.
- **Data Discretization:** Involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals.

SCREENSHOT/OUTPUT:



CONCLUSION:

Thus we learnt to apply preprocessing technique on dataset.

DISCUSSION AND VIVA VOCE:

- What is preprocessing?

  Data preprocessing is a step in the data mining and data analysis process that takes raw data and transforms it into a format that can be understood and analyzed by computers and machine learning.

- Why we need to preprocess the data?

  When using data sets to train machine learning models, you'll often hear the phrase **"garbage in, garbage out"** This means that if you use bad or "dirty" data to train your model, you'll end up with a bad, improperly trained model that won't actually be relevant to your analysis.

- How to perform data cleansing?

Data cleaning is the process of adding missing data and correcting, repairing, or removing incorrect or irrelevant data from a data set. Dating cleaning is the most important step of pre-processing because it will ensure that your data is ready to go for your downstream needs.

Data cleaning will correct all of the inconsistent data you uncovered in your data quality assessment. Depending on the kind of data you're working with, there are a number of possible cleaners you'll need to run your data through.

- Expand .csv?

  Comma-Separated Values (CSV) is a file format used to store tabular data in which numbers and text are stored in a plain-text form that can be easily written and read in a text editor. CSV format is the most common import and export format for spreadsheets and databases.

- Differentiate .arff and .csv?

  ARFF is an acronym that stands for Attribute-Relation File Format. It is an extension of the CSV file format where a header is used that provides metadata about the data types in the columns.

REFERENCE:

- *https://www.iitr.ac.in/media/facspace/patelfec/16Bit/slides/Lecture-2-Data-Preprocessing-Part-1.pdf*
- *www.cs.sfu.ca/~han/DMbook.html*
- *https://publish.pothi.com/preview/?sku=ebook3591*
- *http://www.cs.ccsu.edu/~markov/weka-tutorial.pdf*
- *Data Mining – Concepts and Techniques, Jiawei Han & Micheline Kamber, Morgan Kaufmann Publishers, Elsevier, 2nd Edition, 2006.*

Practical No.4

Aim: Demonstration of Association Rule process on dataset using Apriori Algorithm.

AIM: Demonstration of Association Rule process on dataset using Apriori Algorithm.

OBJECTIVES:

- To analyze various association technique.
- To Experiment with various algorithms available for association.

AIM: Demonstration of Association Rule process on dataset using Apriori Algorithm.

OBJECTIVES:

- To analyze various association technique.
- To Experiment with various algorithms vailable for association.

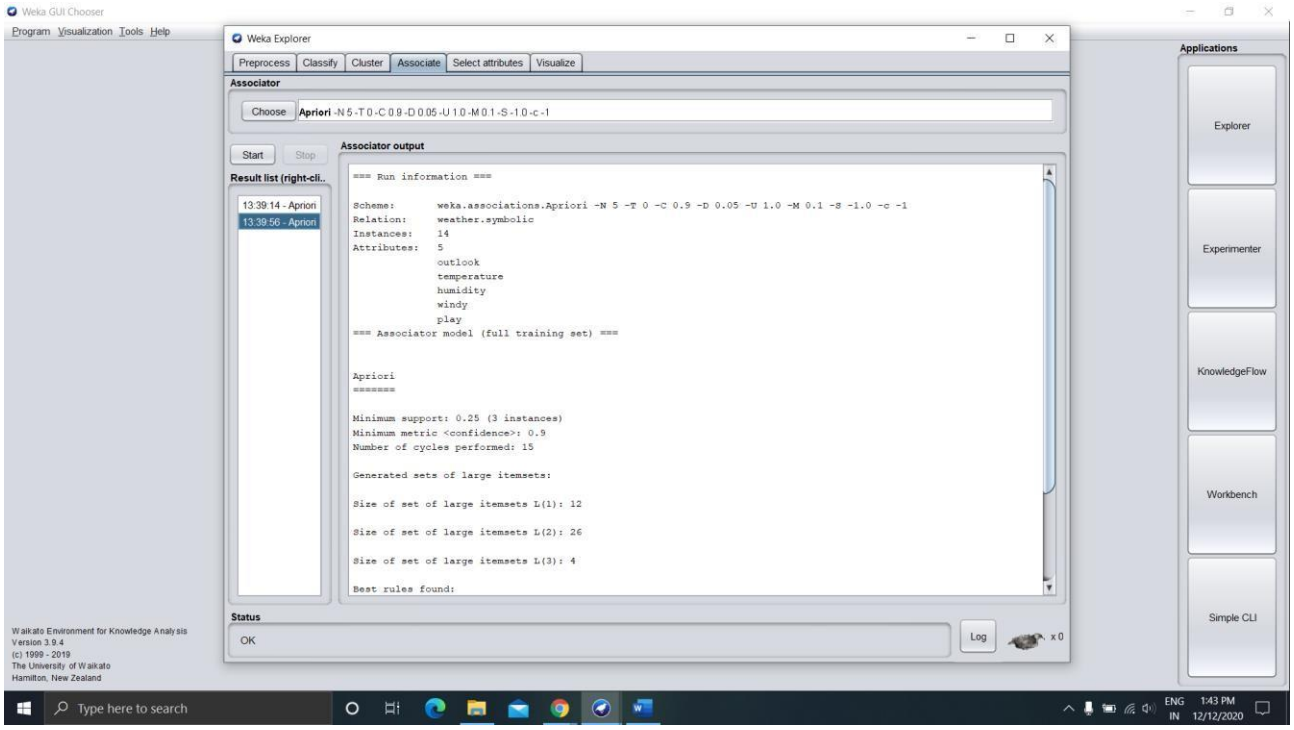
THEORY:

Frequent Item Set (FIS) mining is an essential part of many Machine Learning algorithms. What this technique is intended to do is to extract the most frequent and largest item setswithin a big list of transactions containing several items each. For example:Association rules are created by analyzing data for frequent if/then patterns and using the criteria *support* and *confidence* to identify the most important relationships. *Support* is an indication of how frequently the items appear in the database. *Confidence* indicates the number of times the if/then statements have been found to be true.

In data mining, association rules are useful for analyzing and predicting customer behavior. They play an important part in shopping basket data analysis, product clustering, catalog design and store layout.

ALGORITHM:

OUTPUT:



CONLUSION: Thus we learnt to analyze various association technique using Apriori method.

DISCUSSION AND VIVA VOCE:

- What is frequest item set?

  Support for itemset A: Number of baskets containing all items in A. Given a support threshold s, the set of items that appear in at least s baskets are called frequent itemsets.

- Where you find data association in real life?

  Association Rule Mining is sometimes referred to as "Market Basket Analysis", as it was the first application area of association mining. The aim is to discover associations of items occurring together more often than you'd expect from randomly sampling all the possibilities. The classic anecdote of Beer and Diaper will help in understanding this better.

- Differentiate FP Growth with Apriori

Apriori	FP Growth
Apriori generates the frequent patterns by making the itemsets using pairing such as single item set, double itemset, triple itemset.	FP Growth generates an FP-Tree for making frequent patterns.
Apriori uses candidate generation where frequent subsets are extended one item at a time.	FP-growth generates conditional FP-Tree for every item in the data.
Since apriori scans the database in each of its steps it becomes time-consuming for data where the number of items is larger.	FP-tree requires only one scan of the database in its beginning steps so it consumes less time.
A converted version of the database is saved in the memory	Set of conditional FP-tree for every item is saved in the memory
It uses breadth-first search	It uses a depth-first search.

- Which algorithm is better?

  Both algorithms are good in different application scenarios.

- Explain steps of applying Apriori to dataset.

  Apriori algorithm is a sequence of steps to be followed to find the most frequent itemset in the given database. This data mining technique follows the join and the prune steps iteratively until the most frequent itemset is achieved. A minimum support threshold is given in the problem or it is assumed by the user.

  #1) In the first iteration of the algorithm, each item is taken as a 1-itemsets candidate. The algorithm will count the occurrences of each item.

#2) Let there be some minimum support, min_sup (eg 2). The set of 1 – itemsets whose occurrence is satisfying the min sup are determined. Only those candidates which count more than or equal to min_sup, are taken ahead for the next iteration and the others are pruned.

#3) Next, 2-itemset frequent items with min_sup are discovered. For this in the join step, the 2-itemset is generated by forming a group of 2 by combining items with itself.

#4) The 2-itemset candidates are pruned using min-sup threshold value. Now the table will have 2 –itemsets with min-sup only.

#5) The next iteration will form 3 –itemsets using join and prune step. This iteration will follow antimonotone property where the subsets of 3-itemsets, that is the 2 –itemset subsets of each group fall in min_sup. If all 2-itemset subsets are frequent then the superset will be frequent otherwise it is pruned.

#6) Next step will follow making 4-itemset by joining 3-itemset with itself and pruning if its subset does not meet the min_sup criteria. The algorithm is stopped when the most frequent itemset is achieved.

REFERENCE:

- *http://www.datascienceontology.com/papers/hunyadi.pdf*
- *https://www.worldwidejournals.com/paripex/file.php?val=March_2013_1363611848_829fa_64..pdf*
- *www.cs.sfu.ca/~han/DMbook.html*
- *https://publish.pothi.com/preview/?sku=ebook3591*
- *http://www.cs.ccsu.edu/~markov/weka-tutorial.pdf*
- *Data Mining – Concepts and Techniques, Jiawei Han & Micheline Kamber, Morgan Kaufmann Publishers, Elsevier, 2nd Edition, 2006.*

Practical No.5

Aim : Demonstration of classification rule process on dataset using Naive Baye's Algorithm

AIM: Demonstration of classification rule process on dataset using Naïve Bayes Algorithm

OBJECTIVES:

- To discover the need of data classification.
- To Apply Naive Bayes algorithm to provided dataset and analyze the output.

AIM: Demonstration of classification rule process on dataset using Naïve Bayes Algorithm

OBJECTIVES:

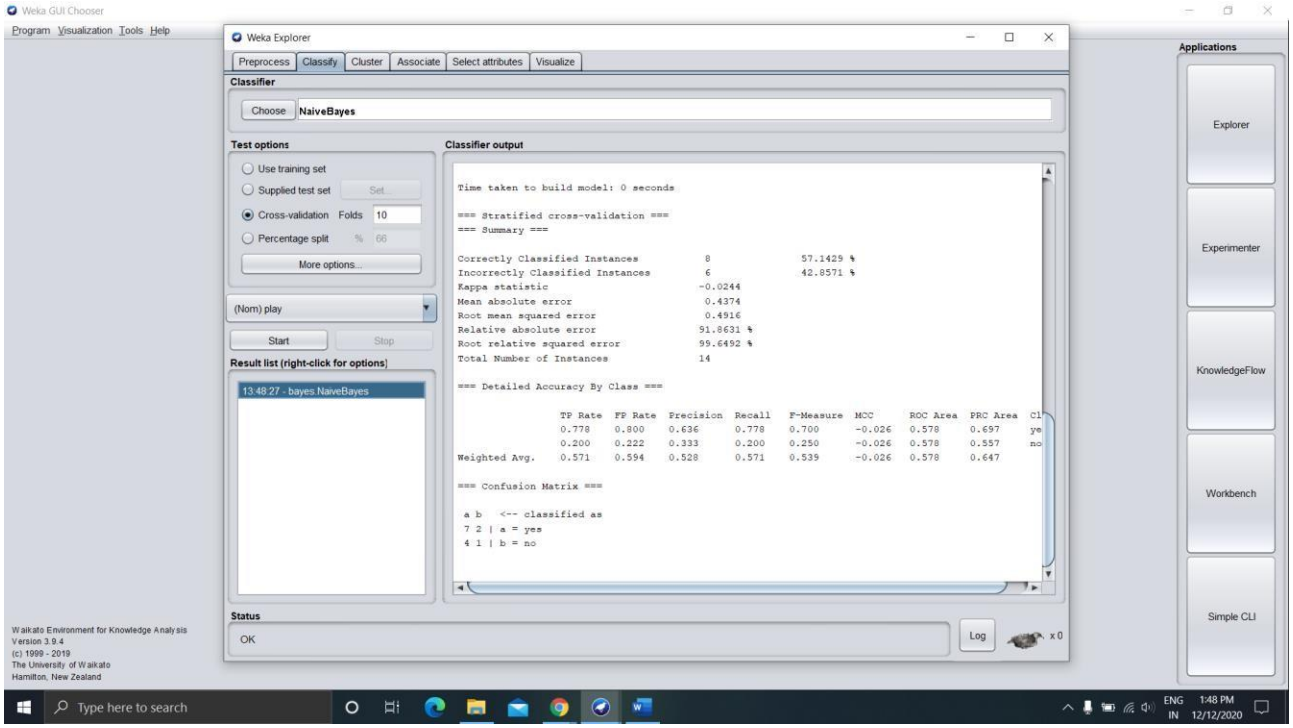
- To discover the need of data classification.
- To Apply Naive Bayes algorithm to provided dataset and analyze the output.

THEORY:

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features. For some types of probability models, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.

ALGORITHM:

OUTPUT:



CONLUSION: Thus, we successfully demonstrated of classification rule process on dataset using naïve bayes Algorithm

DISCUSSION AND VIVA VOCE:

- Why this algorithm is known as naive?

  The Algorithm is ‘naive’ because it makes assumptions that may or may not turn out to be correct.

- Differentiate naive bayes with bayes net algorithm?

  Naive Bayes is a type of prediction model; one which assumes that all of the features are mutually independent.

  The theorem known as “Bayes Theorem” is a theorem. It is a mathematical result. It tells us that $P(A|B)=P(B|A)P(A)/P(B)$

  Bayes Theorem is a theorem that allows us to infer the probability of a particular model given observed data. There are many Machine Learning algorithms based on it. Naive Bayes is one of them.

- Which condition known as naive condition?

  Conditional probability is defined as the likelihood of an event or outcome occurring, based on the occurrence of a previous event or outcome. Conditional probability is calculated by multiplying the probability of the preceding event by the updated probability of the

succeeding, or conditional, event.

- Where we can apply this classifier?

  Naive Bayes algorithms are mostly used in face recognition, weather prediction, Medical Diagnosis, News classification, Sentiment Analysis, etc.

REFERENCE:

http://www.nptel.ac.in/courses/106108057/20

https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm#DMCON004

www.cs.sfu.ca/~han/DMbook.html

https://publish.pothi.com/preview/?sku=ebook3591

http://www.cs.ccsu.edu/~markov/weka-tutorial.pdf

Data Mining – Concepts and Techniques, Jiawei Han & Micheline Kamber, Morgan Kauf-mann Publishers, Elsevier, 2nd Edition, 2006.

Practical No. 6

Aim: Demonstration of classification rule process on dataset using BayesNet Algorithm

AIM: Demonstration of classification rule process on dataset using BayesNet Algorithm

OBJECTIVES:

- To Discover the need of BayesNet Algorithm for dataset processing.
- To Compare the output after applying Naive bayes & BayesNet using visualization technique.

AIM: Demonstration of classification rule process on dataset using BayesNet Algorithm.

OBJECTIVES:

- To Discover the need of BayesNet Algorithm for dataset processing.
- To categorize the data on the basis of anomalies.

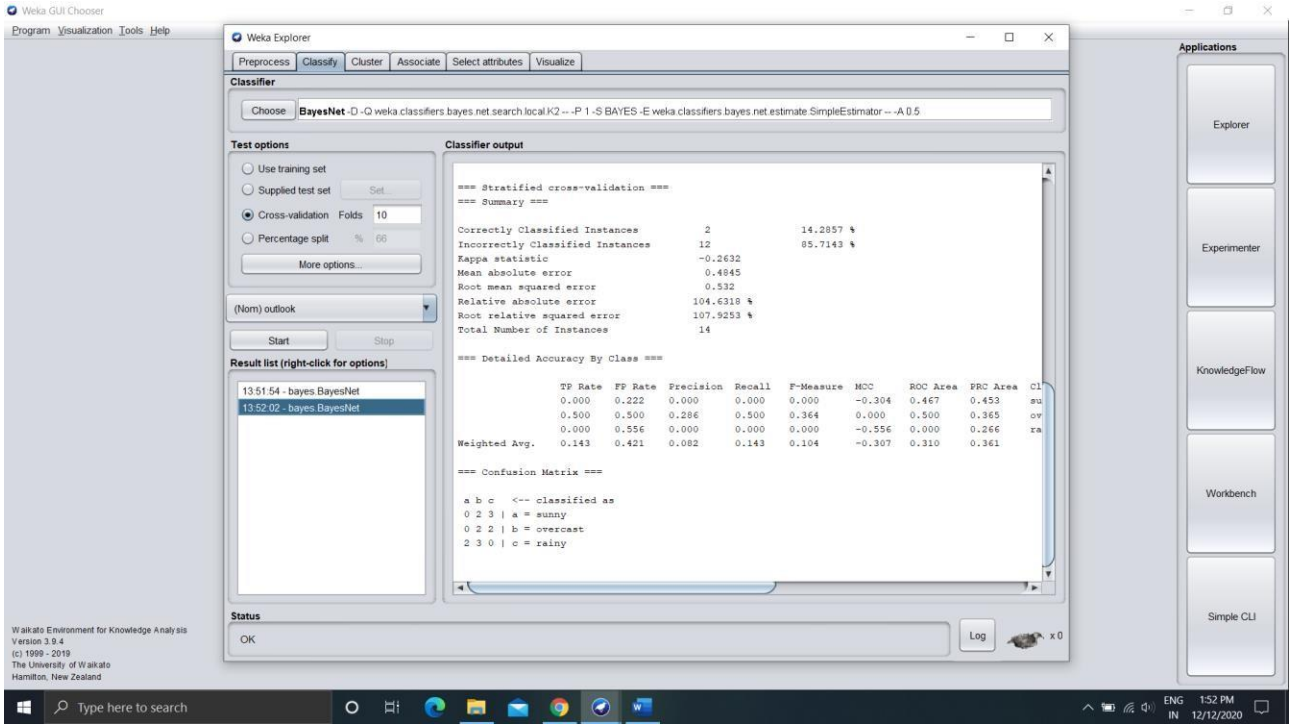
THEORY:

Learning a Bayesian network is a two stage process, a natural division: first learn a network structure, then learn the probability tables.Structure learning is performed via local score metrics, or conditional independence tests or global score metrics.To usea Bayesian network as a classifier, one simply calculates Probability.

The problem with the Naïve Bayes Classifier is that it assumes all attributes are independent of each other which in general can not be applied. Gaussian PDF can be plug-in here to estimate the attribute probability density function (PDF). Because the well developed BayesNet theories, we can classify the new object easier through the same Bayes Classifier Model but with certain degree recognition of the covariance. Normally, this gives more accurate classification result.

ALGORITHM:

OUTPUT:



CONLUSION: Thus we successfully demonstrated of classification rule process on dataset using BayesNet Algorithm.

DISCUSSION AND VIVA VOCE:

- What do you mean by anomaly detection?

  Anomaly detection (aka outlier analysis) is a step-in data mining that identifies data points, events, and/or observations that deviate from a dataset's normal behavior. Anomalous data can indicate critical incidents, such as a technical glitch, or potential opportunities, for instance a change in consumer behavior. Machine learning is progressively being used to auto-mate anomaly detection.

- Differentiate naive bayes with bayes net algorithm?

  Naive Bayes is a type of prediction model; one which assumes that all of the features are mutually independent.

  The theorem known as "Bayes Theorem" is a theorem. It is a mathematical result. It tells us that $P(A|B)=P(B|A)P(A)/P(B)$

  Bayes Theorem is a theorem that allows us to infer the probability of a particular model given observed data. There are many Machine Learning algorithms based on it. Naive Bayes is one of them.

- How to decide outliers in dataset?

Outliers are unusual values in your dataset, and they can distort statistical analyses and violate their assumptions. Unfortunately, all analysts will confront outliers and be forced to make decisions about what to do with them. Given the problems they can cause, you might think that it's best to remove them from your data. But, that's not always the case. Removingoutliers is legitimate only for specific reasons.

- Where we can apply this classifier?

  Naive Bayes algorithms are mostly used in face recognition, weather prediction, Medical Diagnosis, News classification, Sentiment Analysis, etc.

- Explain the need of data classification?

  Data classification has improved significantly over time. Today, the technology is used for a variety of purposes, often in support of data security initiatives. But data may be classified for a number of reasons, including ease of access, maintaining regulatory compliance, and to meet various other business or personal objectives. In some cases, data classification is a regulatory requirement, as data must be searchable and retrievable within specified timeframes. For the purposes of data security, data classification is a useful tactic that facilitates proper security responses based on the type of data being retrieved, transmitted, or copied.

REFERENCE:

http://www.nptel.ac.in/courses/106108057/20

https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm#DMCON004

www.cs.sfu.ca/~han/DMbook.html

https://publish.pothi.com/preview/?sku=ebook3591

http://www.cs.ccsu.edu/~markov/weka-tutorial.pdf

Data Mining – Concepts and Techniques, Jiawei Han & Micheline Kamber, Morgan Kauf-mann Publishers, Elsevier, 2nd Edition, 2006.

Practical No.7

Aim: Demonstration of classification rule process on dataset using j48 Algorithm

AIM: Demonstration of classification rule process on dataset using j48 Algorithm

OBJECTIVES:

- To discover the benefits of using j48 over other bayesian rules.
- To solve classification problem by supervised method (Decision tree) which leads to decisionsupport.

AIM: Demonstration of classification rule process on dataset using j48 Algorithm.

OBJECTIVES:

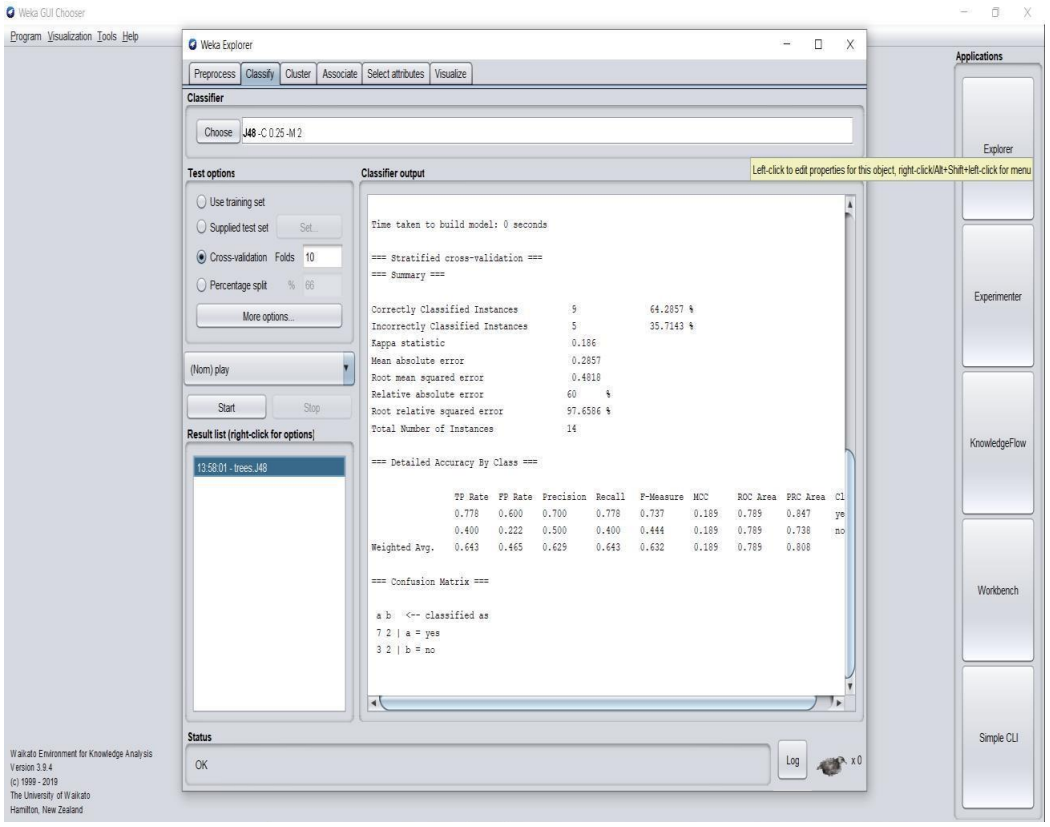
- To analyze the benefits of using j48 over other bayesian rules.
- To solve classification problem by supervised method (Decision Tree) which leads to decision support.

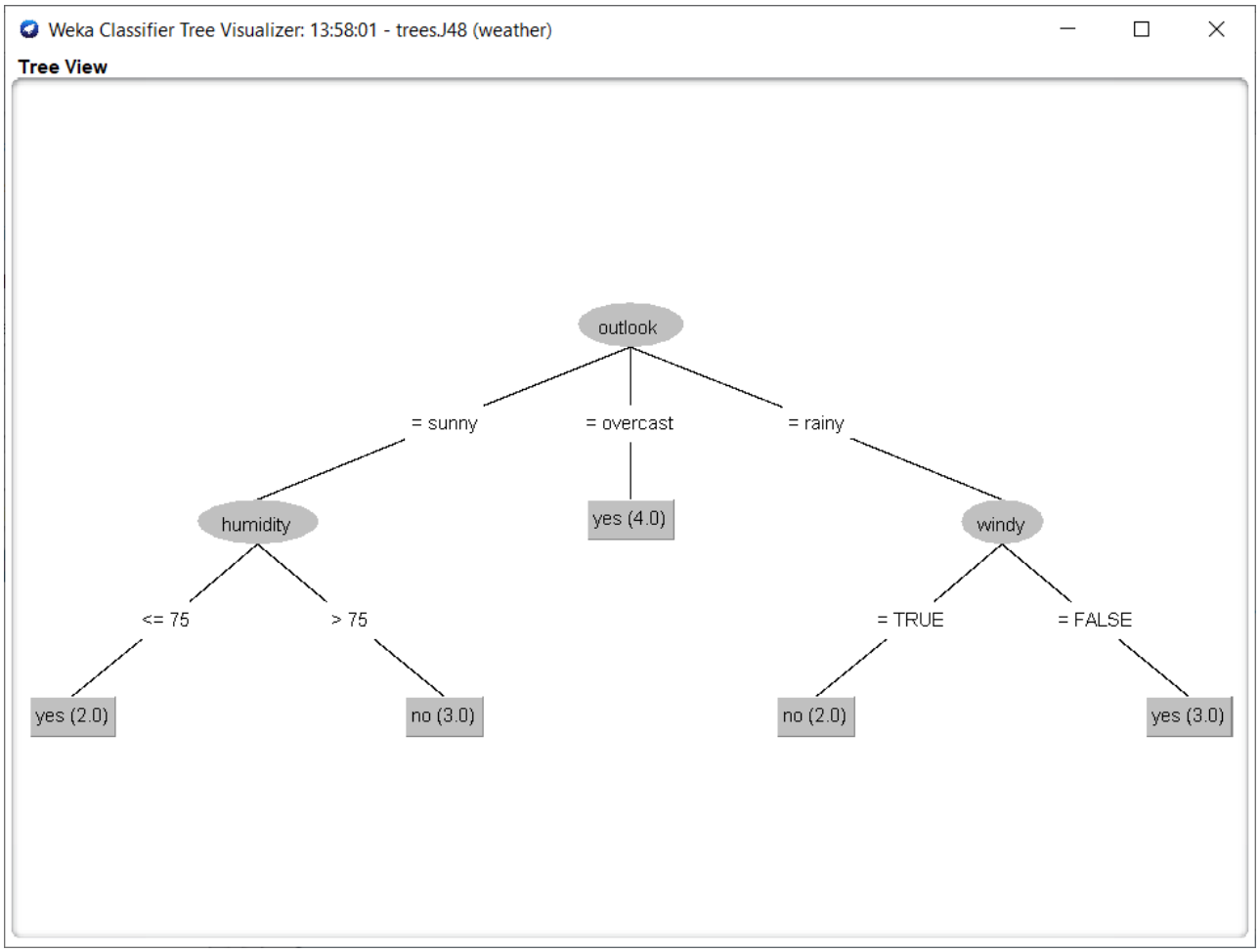
THEORY:

The modified J48 decision tree algorithm examines the normalized information gain that results from choosing an attribute for splitting the data. To make the decision, the attribute with the highest normalized information gain is used. Then the algorithm recurs on the smaller subsets. Thesplitting procedure stops if all instances in a subset belong to the same class.Then a leaf node iscreated in the decision tree telling to choose that class. In this case, the modified J48 decision treealgorithm creates a decision node higher up in the tree using the expected value of the class

ALGORITHM:

OUTPUT:





CONCLUSION: Thus, we successfully demonstrated the classification rule process on dataset using j48 Algorithm

DISCUSSION AND VIVA VOCE:

- Explain working of j48?

  Behind the idea of a decision tree, we will find what it is called `information gain`, a concept that measures the amount of information contained in a set of data. It gives the idea of importance of an attribute in a dataset.

- Which approach is supported by j48?

  Decision tree J48 is the implementation of algorithm ID3 (Iterative Dichotomiser 3) developed by the WEKA project team. R includes this nice work into package RWeka.

- Why to use decision tree?

  By applying a decision tree like J48 on that dataset would allow you to predict the target variable of a new dataset record.

- Where we can apply this classifier?
- How many methods are used for data classification?

REFERENCE:

http://www.nptel.ac.in/courses/106108057/20

https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm#DMCON004

www.cs.sfu.ca/~han/DMbook.html

https://publish.pothi.com/preview/?sku=ebook3591

http://www.cs.ccsu.edu/~markov/weka-tutorial.pdf

Data Mining – Concepts and Techniques, Jiawei Han & Micheline Kamber, Morgan Kauf-mann Publishers, Elsevier, 2nd Edition, 2006.

Practical No.8

Aim: Demonstration of classification rule process on dataset using random tree Algorithm.

AIM: Demonstration of classification rule process on dataset using random tree Algorithm

OBJECTIVES:

- To develop decision tree from random forest strategy
- To averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance.

AIM: Demonstration of classification rule process on dataset using random tree Algorithm.

OBJECTIVES:

- To develop decision tree from random forest strategy
- To averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance.

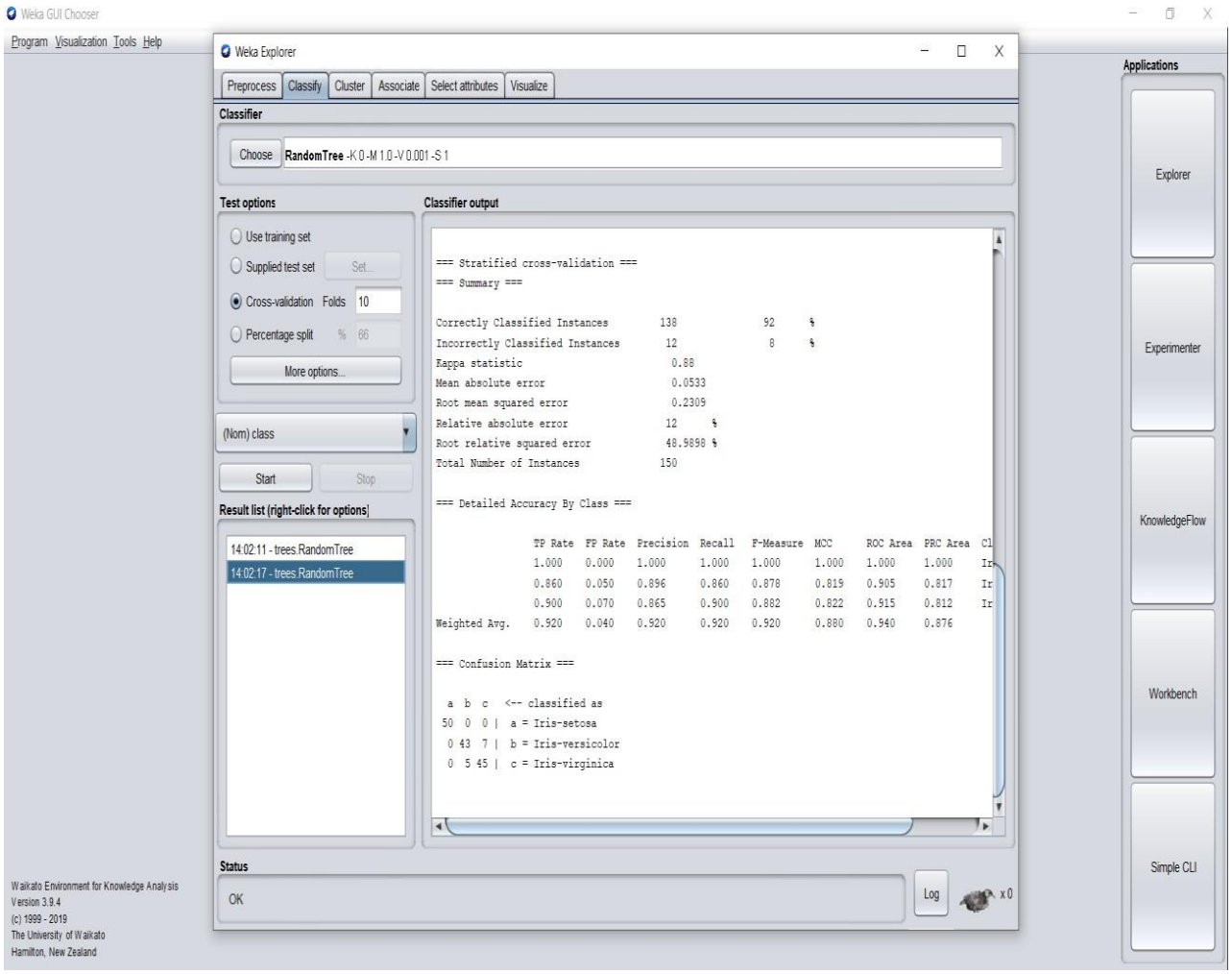
THEORY:

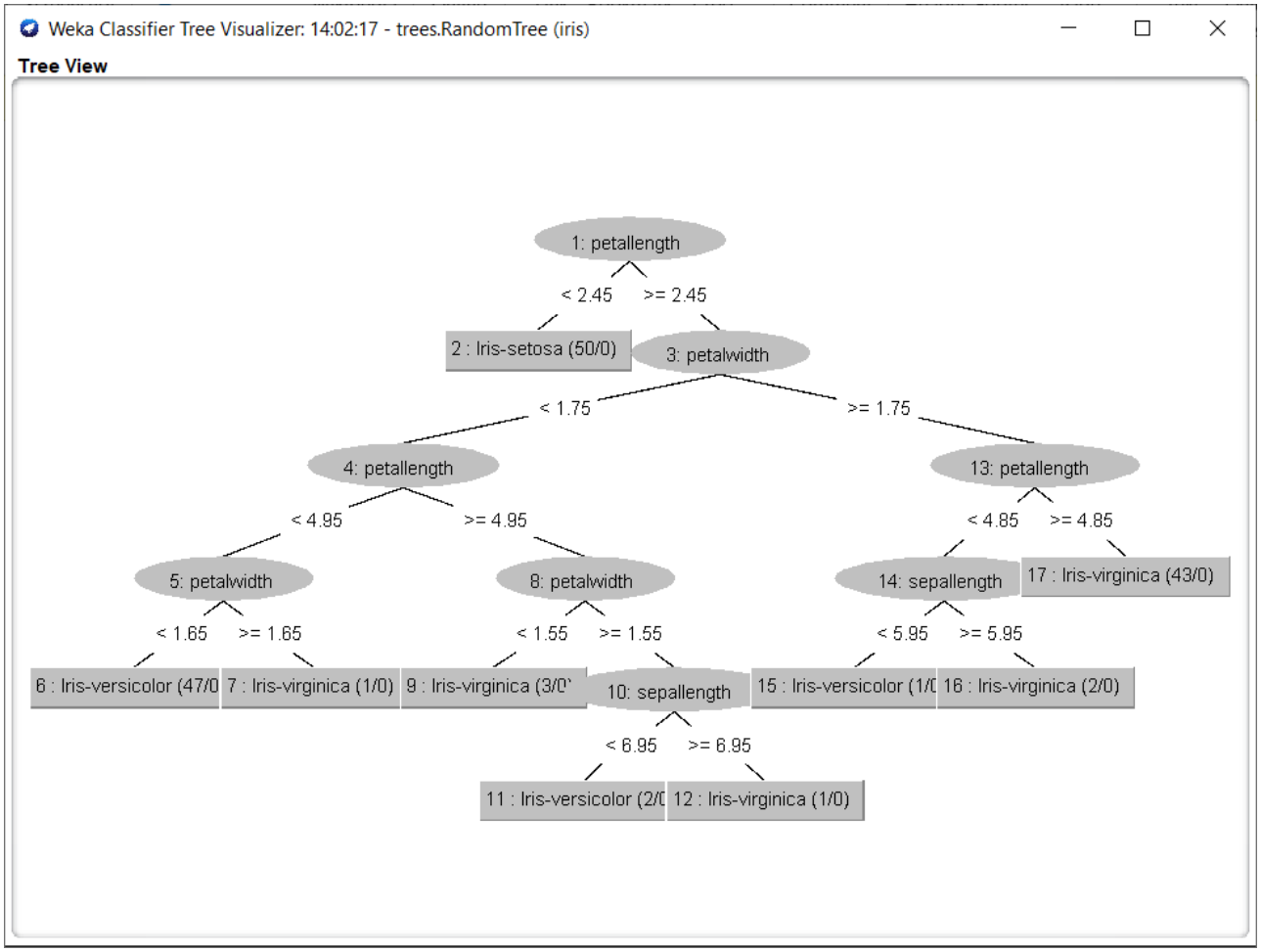
Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

The first algorithm for random decision forests was created by Tin Kam Ho using the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification.

ALGORITHM:

OUTPUT:


Weka GUI Chooser
Program Visualization Tools Help
Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Classifier
Choose RandomTree -K 0 -M 1.0 -V 0.001 -S 1
Test options
Use training set
Supplied test set Set...
Cross-validation Folds 10
Percentage split % 66
More options...
(Nom) class
Start Stop
Result list (right-click for options)
14:02:11 - trees.RandomTree
14:02:17 - trees.RandomTree
Classifier output
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 138 92 %
Incorrectly Classified Instances 12 8 %
Kappa statistic 0.88
Mean absolute error 0.0533
Root mean squared error 0.2309
Relative absolute error 12 %
Root relative squared error 48.9898 %
Total Number of Instances 150
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Cl
1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 Ir
0.860 0.050 0.896 0.860 0.878 0.819 0.905 0.817 Ir
0.900 0.070 0.865 0.900 0.882 0.822 0.915 0.812 Ir
Weighted Avg. 0.920 0.040 0.920 0.920 0.920 0.880 0.940 0.876
=== Confusion Matrix ===
a b c <-- classified as
50 0 0 | a = Iris-setosa
0 43 7 | b = Iris-versicolor
0 5 45 | c = Iris-virginica
Status
OK
Log
x 0
Applications
Explorer
Experimenter
KnowledgeFlow
Workbench
Simple CLI
Waikato Environment for Knowledge Analysis
Version 3.9.4
(c) 1999 - 2019
The University of Waikato
Hamilton, New Zealand



CONLUSION: Thus, we successfully demonstrated the classification rule process on dataset using random tree Algorithm.

DISCUSSION AND VIVA VOCE:

- Why this algorithm is known as random?

Two key concepts that give it the name random:

1. A random sampling of training data set when building trees.
2. Random subsets of features considered when splitting nodes.

- Differentiate all supervised strategies used for classification?

Logistic regression is kind of like linear regression, but is used when the dependent variable is not a number but something else (e.g., a "yes/no" response). It's called regression but performs classification based on the regression and it classifies the dependent variable into either of the classes.

K-NN algorithm is one of the simplest classification algorithms and it is used to identify the data points that are separated into several classes to predict the classification of a new sample point. K-NN is a non-parametric**,** lazy learning algorithm. It classifies new cases based on a similarity measure (i.e., distance functions).

Support vector is used for both regression and classification. It is based on the concept of decision planes that define decision boundaries. A decision plane (hyperplane) is one that separates between a set of objects having different class memberships.

The naive Bayes classifier is based on Bayes' theorem with the independence assumptions between predictors (i.e., it assumes the presence of a feature in a class is unrelated to any other feature). Even if these features depend on each other, or upon the existence of the other features, all of these properties independently. Thus, the name naive Bayes.

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. It follows Iterative Dichotomiser 3(ID3) algorithm structure for determining the split.

- How random tree reduces the variant values?

  In the bagging technique, a data set is divided into **N** samples using randomized sampling. Then, using a single learning algorithm a model is built on all samples. Later, the resultant predictions are combined using voting or averaging in parallel.

- Can we apply random tree on unsupervised way?

  Yes, we can apply random tree on unsupervised way. In the unsupervised case we don't have labels to train on. Instead, like other clustering procedures, need to find the underlying structure in the data.

- Do we discretize the attribute before applying this algorithm?

  Yes, we discretize the attribute before applying this algorithm.

REFERENCE:

http://www.nptel.ac.in/courses/106108057/20

https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm#DMCON004

www.cs.sfu.ca/~han/DMbook.html

https://publish.pothi.com/preview/?sku=ebook3591

http://www.cs.ccsu.edu/~markov/weka-tutorial.pdf

Data Mining – Concepts and Techniques, Jiawei Han & Micheline Kamber, Morgan Kauf-mann Publishers, Elsevier, 2nd Edition, 2006.

Practical No.9

Aim: Demonstration of clustering rule process on dataset using DBSCAN.

AIM: Demonstration of clustering rule process on dataset using DBSCAN.

OBJECTIVES:

- To understand the concept of cluster by analyzing the dataset with random values.
- To differentiate classsification and clustering and find its need.

AIM: Demonstration of clustering rule process on dataset using DBSCAN.

OBJECTIVES:

- To understand the concept of cluster by analyzing the dataset with random values.
- To differentiate classsification and clustering and find its need.

THEORY:

Density based clustering algorithm has played a vital role in finding non linear shapes structure based on the density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is most widely used density based algorithm. It uses the concept of density reachability and density connectivity. On the other hand, K-means (MacQueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assumek clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result.

ALGORITHM:

OUTPUT:

CONCLUSION: Thus, we successfully demonstrated the clustering rule process on dataset using DBSCAN.

DISCUSSION AND VIVA VOCE:

- Expand DBSCAN?

  DBSCAN stands for **Density-Based Spatial Clustering of Applications with Noise**.
- What is clustering?

  The process of combining a set of physical or abstract objects into classes of the same objects is known as clustering. A cluster is a set of data objects that are the same as one another within the same cluster and are disparate from the objects in other clusters. A cluster of data objects can be considered collectively as one group in several applications. Cluster analysis is an essential human activity.
- Why we need to cluster the data?

When it comes to business, data mining is most commonly used by companies with a strong focus on customers – so retail, finance, and marketing are some of the key organisations that benefit from data mining. Data mining is so important to these kinds of businesses because it allows them to ‘drill down’ into the data, and using clustering methods to analyse the data can help them gain further insights from the data they have on file. From this they can examine the relationships between both internal factors – pricing, prod-uct positioning, staff skills – and external factors – such as competition and the demographics of customers. For instance, utilising one of the clustering methods during dat mining can help business to identify distinct groups within their customer base. They can cluster different customer types into one group based on different factors, such as purchasing patterns. The factors analysed through clustering can have a big impact on sales and cus- tomer satisfaction, making it an invaluable tool to boost revenue, cut costs, or sometimes even both.

REFERENCE:

Practical No.10

Aim: Demonstration of clustering rule process on dataset using simple k-means

AIM: Demonstration of clustering rule process on dataset using simple k-means.

OBJECTIVES:

- To understand the concept of k-means algorithm and apply it on dataset.
- To relate k-means with signal processing.

AIM: Demonstration of clustering rule process on dataset using simple k-means.

OBJECTIVES:

- To understand the concept of k-means algorithm and apply it on dataset.
- To relate k-means with signal processing.

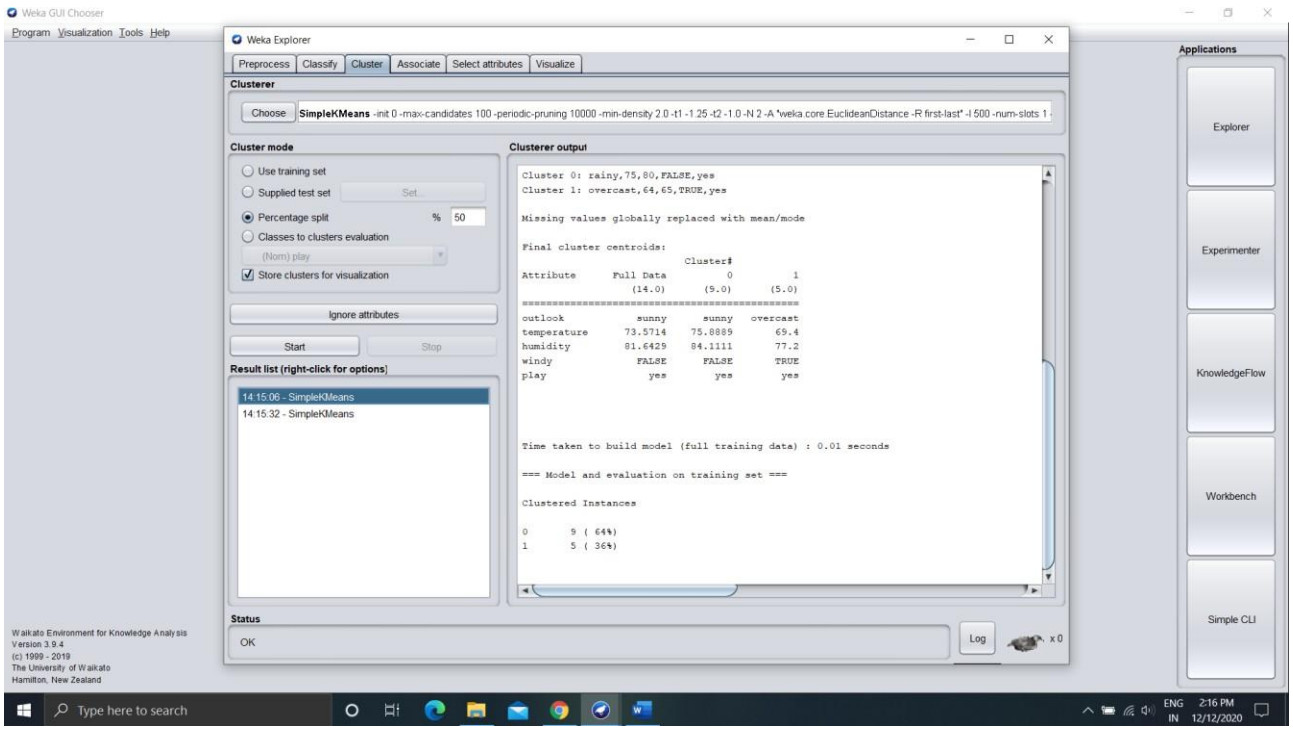
THEORY

K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. *k*-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. The problem is computationally difficult (NP-hard); however, there are efficient heuristic algorithms that are commonly employed and converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both algorithms. Additionally, they both use cluster centers to model the data; however, *k*-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

The algorithm has a loose relationship to the *k*-nearest neighbor classifier, a popular machine learning technique for classification that is often confused with *k*-means because of the *k* in the name. One can apply the 1-nearest neighbor classifier on the cluster centers obtained by *k*-means to classify new data into the existing clusters. This is known as nearest centroid classifier or Rocchio algorithm.

ALGORITHM:

OUTPUT:



CONCLUSION: Thus, we successfully demonstrated the clustering rule process on dataset using simple k-means.

DISCUSSION AND VIVA VOCE:

- K means works on which strategy?

  K-means algorithm is an iterative algorithm that tries to partition the dataset into *K*pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to **only one group**. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

- Differentiate all k-means with DBSCAN?

S.No.	K-means Clustering	DBScan Clustering
1.	Clusters formed are more or less spherical or convex in shape and must have same feature size.	Clusters formed are arbitrary in shape and may not have same feature size.
2.	K-means clustering is sensitive to the number of clusters specified.	Number of clusters need not be specified.

3.	K-means Clustering is more efficient for large datasets.	DBSCan Clustering can not efficiently handle high dimensional datasets.
4.	K-means Clustering does not work well with outliers and noisy datasets.	DBScan clustering efficiently handles outliers and noisy datasets.
5.	In the domain of anomaly detection, this algorithm causes problems as anomalous points will be assigned to the same cluster as "normal" data points.	DBScan algorithm, on the other hand, locates regions of high density that are separated from one another by regions of low density.
6.	It requires one parameter: Number of clusters (**K**)	It requires two parameters: Radius(**R**) and Minimum Points(**M**) R determines a chosen radius such that if it includes enough points within it, it is a dense area. M determines the minimum number of data points required in a neighborhood to be defined as a cluster.
7.	Varying densities of the data points doesn't affect K-means clustering algorithm.	DBScan clustering does not work very well for sparse datasets or for data points with varying density.

- What is cluster?

  A cluster is a set of data objects that are the same as one another within the same cluster and are disparate from the objects in other clusters. A cluster of data objects can be considered collectively as one group in several applications. Cluster analysis is an essential human activity.

- Difference between classification & clustering?

Classification

- It is used with supervised learning.
- It is a process where the input instances are classified based on their respective class labels.
- It has labels hence there is a need to train and test the dataset to verify the model.
- It is more complex in comparison to clustering.
- Examples: Logistic regression, Naive Bayes classifier, Support vector machines.

Clustering

- It is used with unsupervised learning.
- It groups the instances based on how similar they are, without using class labels.
- It is not needed to train and test the dataset.

- It is less complex in comparison to classification.
- Examples: k-means clustering algorithm, Gaussian (EM) clustering algorithm.

 Which should be applied first on dataset clustering or classification?

Generally, you want your training and validation data sets be separate as much as possible. Ideally, the validation set data would have been obtained only after the model has been trained. If you perform dimensionality reduction before splitting your data to separate sets, you break this isolation between the training and the validation and you won't be sure whether the dimensionality reduction process was over-fitted until your model is tested in real life.

Having said that, there are cases, where efficient separation to training, testing and validation sets is not feasible and other sampling techniques, such as cross validation, leave k out etc are used. In these cases, reducing the dimensionality before the sampling might be the right approach.

REFERENCE:

Practical No.11

Aim: Introduction to new mining tool rapidminer.

AIM: Introduction to new mining tool rapidminer.

OBJECTIVES:

- To understand the functionality of rapidminer.
- Introduction to new mining tool rapidminer, Compare rapidminer with WEKA.
- To check the efficiency for various methods of data mining.

AIM: Introduction to new mining tool rapidminer.

OBJECTIVES:

- To understand the functionality of rapidminer.
- Introduction to new mining tool rapidminer, Compare rapidminer with WEKA.
- To check the efficiency for various methods of data mining.

THEORY:

Data Mining with WEKA

WEKA is java based open source data mining tool which has collection of data mining algorithms such as lazy, rules, decision trees and so on. WEKA opens with 4 options (Explorer, Experimenter, KnowledgeFlow and Simple CLI). Mainly, Explorer and Experimenter are used for data mining. For multiple algorithms comparison, Experimenter is used but for specific results of data mining, Explorer is used. Explorer opens with a screen of data preprocessing. The difficulty on WEKA is opening file because most of data sets are in excel and excel can turn into CSV format but excel file is semicolon but CSV must be for comma, so it needs to convert in text file format but it takes time. However; information on algorithms (capabilities and descriptions) and user options are the best features about WEKA, especially any user can use without any training as well as user can implement its own algorithm.

Data Mining with Rapidminer

RapidMiner is another data mining tool which has ARENA simulation like environment. because every process is described in a similar manner. RapidMiner can use every algorithm in WEKA as well as its own algorithm and R programming can be open in RapidMiner with a connection like any other algorithm. Maybe, its difficulty is that's not easy to use and the result is only based on confusion matrix.

ALGORITHM: If Applicable

OUTPUT:

CONCLUSION:

DISCUSSION AND VIVA VOCE:

- Which tool gives better result?
- How many file formats supported by rapidminer?
- Explain installation process of rapidminer?

REFERENCE: