ANSWER KEY

Data Mining with Business Intelligence
20MCA252

Question-1

What is Regression? Explain Logistic regression with an example. OR

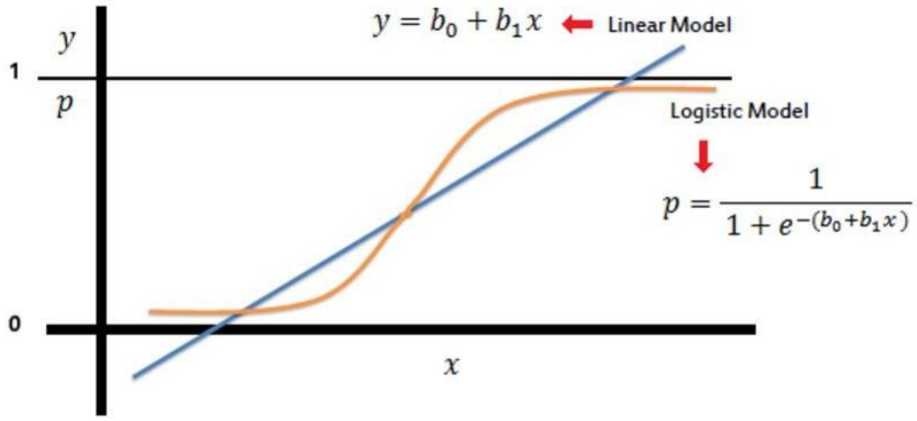
- Regression is a data mining function that predicts a number.
- Age, weight, distance, temperature, income, or sales could all be predicted using regression techniques.
- For example, a regression model could be used to predict children's height, given their age, weight, and other factors.
- A regression task begins with a data set in which the target values are known.
- For example, a regression model that predicts children's height could be developed based on observed data for many children over a period of time.
- The data might track age, height, weight, developmental milestones, family history, and so on.
- Height would be the target, the other attributes would be the predictors, and the data for each child would constitute a case.
- Regression models are tested by computing various statistics that measure the difference between the predicted values and the expected values.
- It is required to understand the mathematics used in regression analysis to develop quality regression models for datamining.
- The goal of regression analysis is to determine the values of parameters for a function that cause the function to best fit a set of data observations that you provide.
- It shows that regression is the process of estimating the value of a continuous target (y) as a function (F) of one or more predictors (x1 , x2 , ..., xn), a set of parameters (θ1 , θ2 , ..., θn), and a measure of error (e).

$$\mathbf{y = F(x,\theta) + e}$$

Logistic Regression

- Logistic regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy).
- The prediction is based on the use of one or several predictors (numerical and categorical).

- A linear regression is not appropriate for predicting the value of a binary variable for two reasons:
  - A linear regression will predict values outside the acceptable range (e.g. predicting probabilities outside the range 0 to 1).
  - Since the dichotomous experiments can only have one of two possible values for each experiment, the residuals will not be normally distributed about the predicted line.
- A logistic regression produces a logistic curve, which is limited to values between 0 and 1.
- Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the “odds” of the target variable, rather than the probability.
- Moreover, the predictors do not have to be normally distributed or have equal variance in each group.



- In the logistic regression the constant (b0) moves the curve left and right and the slope (b1) defines the steepness of the curve.

- Advantage of logistic regression is that the algorithm is highly flexible, taking any kind of input, and supports several different analytical tasks:
    - Use demographics to make predictions about outcomes, such as risk for a certain disease.
    - Explore and weight the factors that contribute to a result. For example, find the factors that influence customers to make a repeat visit to a store.
    - Classify documents, e-mail, or other objects that have many attributes.

Question-2
Write a note on rule based classification.

- Rule-based classifier makes use of a set of IF-THEN rules for classification.
- We can express a rule in the following from

IF *condition* THEN *conclusion*

- Let us consider a rule R1,

R1: **IF age=youth AND student=yes**
THEN buy_computer=yes

- The IF part of the rule is called rule antecedent or precondition.
- The THEN part of the rule is called rule consequent.
- The antecedent part the condition consist of one or more attribute tests and these tests are logically ANDed.
- The consequent part consists of class prediction.
- We can also write rule R1 as follows:

R1: (age = youth) ^ (student = yes))(buys_computer = yes)

- If the condition (that is, all of the attribute tests) in a rule antecedent holds true for a given tuple, we say that the rule antecedent is satisfied (or simply, that the rule is satisfied) and that the rule covers the tuple.

- A rule R can be assessed by its coverage and accuracy.
- Given a tuple, X, from a class labeled data set D, let ncovers be the number of tuples covered by R; ncorrect be the number of tuples correctly classified by R; and ❙D❙ be the number of tuples in D.
- We can define the coverage and accuracy of R as

$$coverage(R) = \frac{n_{covers}}{|D|}$$

$$accuracy(R) = \frac{n_{correct}}{n_{covers}}.$$

- That is, a rule's coverage is the percentage of tuples that are covered by the rule (i.e. whose attribute values hold true for the rule's antecedent).
- For a rule's accuracy, we look at the tuples that it covers and see what percentage of them the rule can
  correctly classify.
- We can use rule-based classification to predict the class label of a given tuple X.
- If a rule is satisfied by X, the rule is said to be triggered.
- For example, suppose we have

X= (age = youth, income = medium, student = yes, credit rating = fair)

- We would like to classify X according to buys_computer. X satisfies R1, which triggers the rule.
- If R1 is the only rule satisfied, then the rule fires by returning the class prediction for X.
- If more than one rule is triggered, we need a conflict resolution strategy to figure out which rule gets to fire and assign its class prediction to X.
- There are many possible strategies. We look at two, namely **size ordering** and **rule ordering**.
- **Size ordering**
  - The size ordering scheme assigns the highest priority to the triggering rule that has the "toughest"
    requirements, where toughness is measured by the rule antecedent size.
  - That is, the triggering rule with the most attribute tests is fired.
- **Rule ordering**
  - The rule ordering scheme prioritizes the rules beforehand. The ordering may be class based or rule- based.
  - With class-based ordering, the classes are sorted in order of decreasing "importance," such as by
    decreasing order of prevalence.
  - That is, all of the rules for the most prevalent (or most frequent) class come first, the rules for the next prevalent class come next, and so on.
  - With rule-based ordering, the rules are organized into one long priority list, according to some measure of rule quality such as accuracy, coverage, or size (number of attribute tests in the rule antecedent), or based on advice from domain experts.
  - When rule ordering is used, the rule set is known as a decision list.
  - With rule ordering, the triggering rule that appears earliest in the list has highest priority, and so it gets to fire its class prediction.
  - Any other rule that satisfies X is ignored. Most rule-based classification systems use a class-based rule-ordering strategy.

Question-3
Write ID3 algorithm. Explain with example. OR

ENTROPY MEASURES HOMOGENEITY OF EXAMPLES

- S is a collection of training examples
  - p_+ the proportion of positive examples in $S\ \left(\frac{P}{P+N}\right)$
  - p_- the proportion of negative examples in $S\ \left(\frac{N}{P+N}\right)$
- ***Entropy is 0 if all members of S belong to the same class.(all are either +ve or –ve)***
- ***Entropy is 1 when the collection contains equal number of +ve and –ve examples.***
- ***If the collection contains unequal number of +ve and –ve examples, entropy is between 0 and 1.***

Design the decision tree for the following dataset and predict whether Golf will be played on the day.

Day	Outlook	Temperature	Humidity	Wind	Play Golf
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Attribute: Outlook

$Values\,(Outlook)\ =\ Sunny, Overcast, Rain$

$S\ =\ [9+, 5\ -]$ $\qquad Entropy(S) = -\frac{9}{14}log_2\frac{9}{14} - \frac{5}{14}log_2\frac{5}{14} = 0.94$

$S_{Sunny} \leftarrow [2+, 3-]$ $\qquad Entropy(S_{Sunny}) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.971$

$S_{Overcast} \leftarrow [4+, 0-]$ $\qquad Entropy(S_{Overcast}) = -\frac{4}{4}log_2\frac{4}{4} - \frac{0}{4}log_2\frac{0}{4} = 0$

$S_{Rain}\ \leftarrow [3+, 2-]$ $\qquad Entropy(S_{Rain}) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.971$

$$Gain\ (S, Outlook) = Entropy\ (S) - \sum_{v\,\in\{Sunny, Overcast, Rain\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Outlook)$$

$$= Entropy(S) - \frac{5}{14}Entropy(S_{Sunny}) - \frac{4}{14}Entropy\ (S_{Overcast}\ )$$

$$-\frac{5}{14}Entropy(S_{Rain})$$

$$Gain(S, Outlook) = 0.94 - \frac{5}{14}0.971 - \frac{4}{14}0 - \frac{5}{14}0.971 = 0.2464$$

Attribute: Temp

$Values\ (Temp\ ) = Hot, Mild, Cool$

$S\ =\ [9+, 5\ -]$ $\qquad Entropy(S) = -\frac{9}{14}log\ _2\frac{9}{14} - \frac{5}{14}log\ _2\frac{5}{14} = 0.94$

$S_{Hot} \leftarrow [2+, 2-]$ $\qquad Entropy(S_{Hot}) = -\frac{2}{4}log_2\frac{2}{4} - \frac{2}{4}log_2\ \frac{2}{4} = 1.0$

$S_{Mild} \leftarrow [4+, 2-]$ $\qquad Entropy(S_{Mild}) = -\frac{4}{6}log_2\frac{4}{6} - \frac{2}{6}log_2\frac{2}{6} = 0.9183$

$S_{Cool}\ \leftarrow [3+, 1-]$ $\qquad Entropy(S_{Cool}) = -\frac{3}{4}log_2\frac{3}{4} - \frac{1}{4}log_2\frac{1}{4} = 0.8113$

$$Gain\ (S, Temp) = Entropy(S) - \sum_{v\,\in\{Hot, Mild, Cool\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Temp)$$

$$= Entropy(S) - \frac{4}{14}Entropy(S_{Hot}) - \frac{6}{14}Entropy(S_{Mild})$$

$$-\frac{4}{14}Entropy(S_{Cool})$$

$$Gain(S, Temp) = 0.94 - \frac{4}{14}1.0 - \frac{6}{14}\ 0.9183 - \frac{4}{14}0.8113 = 0.0289$$

Attribute: Humidity

$Values\ (Humidity\ ) = High, Normal$

$S\ =\ [9+, 5\ -]$ $\qquad Entropy(S) = -\frac{9}{14}log\ _2\frac{9}{14} - \frac{5}{14}log\ _2\frac{5}{14} = 0.94$

$S_{High} \leftarrow [3+, 4-]$ $\qquad Entropy(S_{High}) = -\frac{3}{7}log_2\frac{3}{7} - \frac{4}{7}log_2\ \frac{4}{7} = 0.9852$

$S_{Normal} \leftarrow [6+, 1-]$ $\qquad Entropy(S_{Normal}) = -\frac{6}{7}log_2\frac{6}{7} - \frac{1}{7}log_2\frac{1}{7} = 0.5916$

$$Gain\ (S, Humidity) = Entropy(S) - \sum_{v\ \in\{High, Normal\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Humidity)$$

$$= Entropy(S) - \frac{7}{14}Entropy(S_{High}) - \frac{7}{14}Entropy(S_{Normal})$$

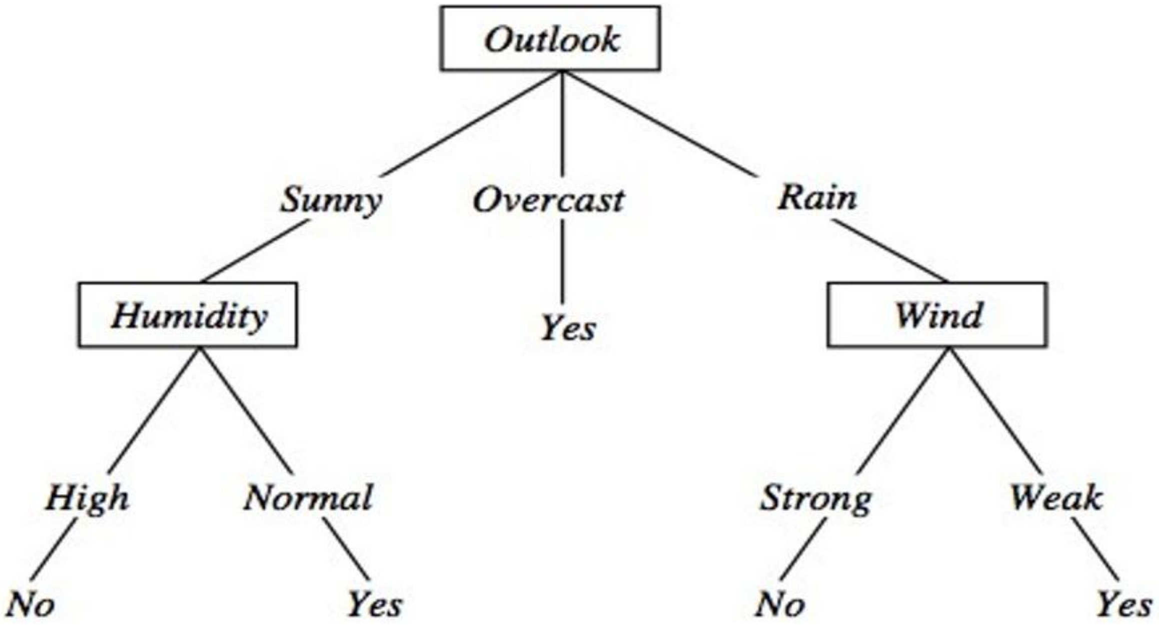
$$Gain(S, Humidity) = 0.94 - \frac{7}{14}0.9852 - \frac{7}{14}\ 0.5916 = 0.1516$$

$Gain\ (S, Outlook\ ) = 0.2464$

$Gain(S, Temp) = 0.0289$

$Gain(S, Humidity) = 0.1516$

$Gain(S, Wind) = 0.0478$



Question-4

The following table gives data set about play tennis. Apply naïve bayes classifier classify the new data (Outlook = Sunny, Temperature = Cool, Humidity = High, Wind = Strong).

Step 1: Convert the data into a frequency table.

Play	Frequency
Yes	9
No	5

Outlook	Yes	No
Sunny	2	3
Overcast	4	0
Rain	3	2

Temperature	Yes	No
Hot	2	2
Mild	4	2
Cool	3	1

Humidity	Yes	No
High	3	4
Normal	6	1

Wind	Yes	No
Strong	3	3
Weak	6	2

Step 2: Create Likelihood table

Play	Frequency	Likelihood
Yes	9	9/14
No	5	5/14

Outlook	Yes	No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Yes	No
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Yes	No
High	3/9	4/5
Normal	6/9	1/5

Wind	Yes	No
Strong	3/9	3/5
Weak	6/9	2/5

Step 3:
New instance
X= (Outlook = Sunny, Temperature = Cool, Humidity = High, Wind = Strong)
Play Tennis = ?

P(X |Play = Yes) = P(Play = Yes) * P(Outlook = Sunny |Yes) * P(Temperature = Cool|Yes) * P(Humidity = High|Yes) * P(Wind = Strong|Yes)

= 9/14 * 2/9 * 3/9 * 3/9 * 3/9
= 0.0053

P(X|Play = No) = P(Play = No) * P(Outlook = Sunny |No) * P(Temperature = Cool|No) * P(Humidity = High|No) * P(Wind = Strong|No)

= 5/14 * 3/5 * 1/5 * 4/5 * 3/5
= 0.0206

So: 0.0206 > 0.0053
Result : X : PlayTennis = No

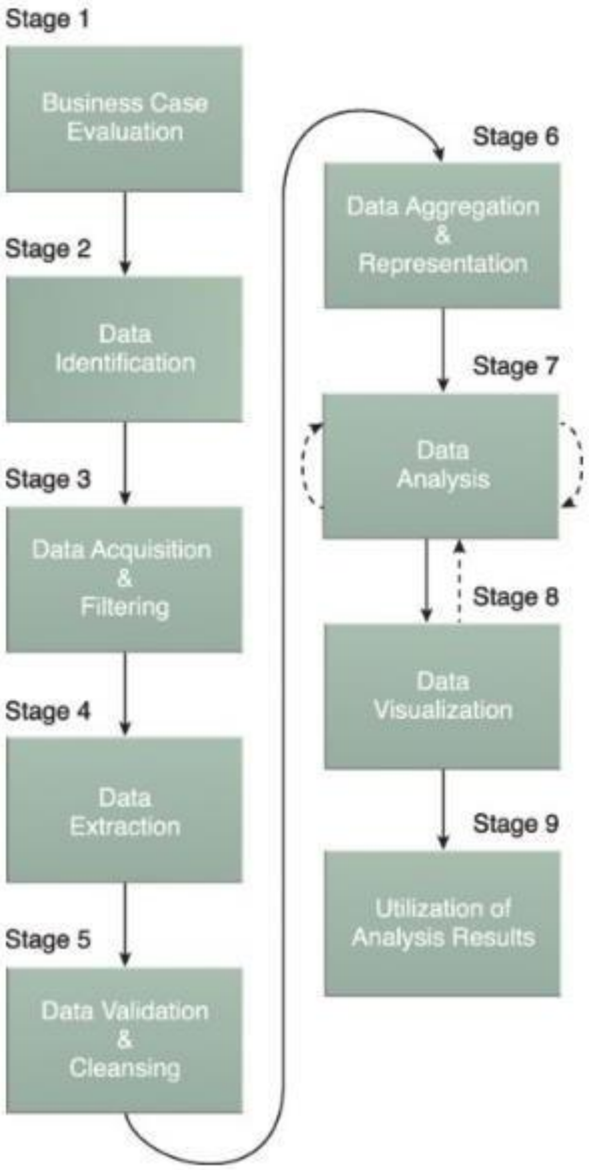
Question-5
Briefly outline the major steps of decision tree classification. OR

Question-6
Explain CART algorithm with error value and Gini Index.
Question-7
What is a neural network? Explain perceptron with algorithm.

and example. OR
Question-8
What is an activation function? Explain the forward pass in a Multilayer neural network.

with an example.
Question-9
Explain Data Analytics Life Cycle in detail . OR

1)



Data Analytics life cycle

- Big Data analysis differs from traditional data analysis primarily due to the volume, velocity and variety characteristics of the data being processes.
- To address the distinct requirements for performing analysis on Big Data, a step-by-step methodology is needed to organize the activities and tasks involved with acquiring, processing, analyzing and repurposing data.
- The upcoming sections explore a specific data analytics lifecycle that organizes and manages the tasks and activities associated with the analysis of Big Data.
- From a Big Data adoption and planning perspective, it is important that in addition to the lifecycle, consideration be made for issues of training, education, tooling and staffing of a data analytics team.
- The Big Data analytics lifecycle can be divided into the following nine stages,

1. Business Case Evaluation
2. Data Identification

3. Data Acquisition & Filtering
4. Data Extraction
5. Data Validation & Cleansing
6. Data Aggregation & Representation
7. Data Analysis
8. Data Visualization
9. Utilization of Analysis Results

Business Case Evaluation

- Each Big Data analytics lifecycle must begin with a well-defined business case that presents a clear understanding of the justification, motivation and goals of carrying out the analysis.
- The Business Case Evaluation stage requires that a business case be created, assessed and approved prior to proceeding with the actual hands-on analysis tasks.
- An evaluation of a Big Data analytics business case helps decision-makers understand the business resources that will need to be utilized and which business challenges the analysis will tackle.
- Based on business requirements that are documented in the business case, it can be determined whether the business problems being addressed are really Big Data problems.
- In order to qualify as a Big Data problem, a business problem needs to be directly related to one or more of the Big Data characteristics of volume, velocity, or variety.

Data Identification

- The Data Identification stage is dedicated to identifying the datasets required for the analysis project and their sources.
- Identifying a wider variety of data sources may increase the probability of finding hidden patterns and correlations. For example, to provide insight, it can be beneficial to identify as many types of related data sources as possible, especially when it is unclear exactly what to look for.
- Depending on the business scope of the analysis project and nature of the business problems being addressed, the required datasets and their sources can be internal and/or external to the enterprise.

Data Acquisition and Filtering

- During the Data Acquisition and Filtering stage, the data is gathered from all of the data sources that were identified during the previous stage.
- The acquired data is then subjected to automated filtering for the removal of corrupt data or data that has been deemed to have no value to the analysis objectives.
- Depending on the type of data source, data may come as a collection of files, such as data purchased from a third-party data provider, or may require API integration, such as with Twitter.
- In many cases, especially where external, unstructured data is concerned, some or most of the acquired data may be irrelevant (noise) and can be discarded as part of the filtering process.

Data Extraction

- The Data Extraction lifecycle stage is dedicated to extracting disparate data and transforming it into a format that the underlying Big Data solution can use for the purpose of the data analysis.
- The extent of extraction and transformation required depends on the types of analytics and capabilities of the Big Data solution.
- For example, extracting the required fields from delimited textual data, such as with webserver log files, may not be necessary if the underlying Big Data solution can already directly process those files.

Data Validation and Cleansing

- The Data Validation and Cleansing stage is dedicated to establishing often complex validation rules and removing any known invalid data.
- Big Data solutions often receive redundant data across different datasets.
- This redundancy can be exploited to explore interconnected datasets in order to assemble validation parameters and fill in missing valid data.

Data Aggregation and Representation

- The Data Aggregation and Representation stage is dedicated to integrating multiple datasets together to arrive at a unified view.
- Performing this stage can become complicated because of differences in:
    - Data Structure – Although the data format may be the same, the data model maybe different.
    - Semantics–A value that is labeled differently in two different datasets may mean the same
      thing, for example "surname" and "last name."
- The large volumes processed by Big Data solutions can make data aggregation a time and effort- intensive operation.
- Reconciling these differences can require complex logic that is executed automatically without the need for human intervention.

Data Analysis

- The Data Analysis stage is dedicated to carrying out the actual analysis task, which typically involves one or more types of analytics.
- This stage can be iterative in nature, especially if the data analysis is exploratory, in which case analysis is repeated until the appropriate pattern or correlation is uncovered.
- The exploratory analysis approach will be explained shortly, along with confirmatory analysis.

Data Visualization

- The Data Visualization stage is dedicated to using data visualization techniques and tools to graphically communicate the analysis results for effective interpretation by business users.
- The results of completing the Data Visualization stage provide users with the ability to perform visual analysis, allowing for the discovery of answers to questions that users

havenotyetevenformulated.

Utilization of Analysis Results

- The Utilization of Analysis Results stage is dedicated to determining how and where processedanalysis data can be further leveraged.
- Depending on the nature of the analysis problems being addressed, it is possible for the analysis results to produce “models” that encapsulate new insights and understandings about the nature of the patterns and relationships that exist within the data that was analyzed.

Question-10
Write a short note on DB Miner /WEKA/DTREG Tools.

1) DB Miner

- DBMiner, a data mining system for interactive mining of multiple-level knowledge in large relational databases, has been developed based on our years-of-research.
- The system implements a wide spectrum of data mining functions, including generalization, characterization, discrimination, association, classification, and prediction.
- By incorporation of several interesting data mining techniques, including attribute-oriented induction, progressive deepening for mining multiple-level rules, and meta-rule guided knowledge mining, the system provides a user-friendly, interactive data mining environment with good performance.

WEKA

- Weka is a collection of machine learning algorithms for data mining tasks.
- The algorithms can either be applied directly to a dataset or called from your own Java code.
- Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.
- It is also well-suited for developing new machine learning schemes.

DTREG

- It is a robust application that is installed easily on any Windows system.
- DTREGreadsCommaSeparatedValue(CSV)datafilesthatareeasilycreatedfromalmostanydata source. Once you create your data file, just feed it into DTREG, and let DTREG do all of the work of creating a decision tree, Support Vector Machine, K-Means clustering, Linear Discriminant Function, Linear Regression or Logistic Regression model. Even complex analysescanbesetupinminutes.
- Classification and Regression Trees. DTREG can build Classification Trees where the target variable being predicted is categorical and Regression Trees where the target variable is continuouslikeincomeorsales volume.