

# **DATA MINING PROJECT REPORT ON BIKE SHARING CATEGORIZATION**



**Submitted By**

14MI503 - Abhishek

14MI508 - Rishabh

**Department of Computer Science and Engineering  
National Institute of Technology Hamirpur  
April, 2018**

## INDEX

S. no.	Topic	Page no.
1.	Introduction	3
2.	Attribute information	3
3.	Technology used	5
4.	Decision tree	5
5.	Construction of decision tree	5
6.	Advantages	5
7.	Disadvantages	5
8.	Explanation	6
9.	Result and applications	7
10.	Screenshots	8

## FIGURE INDEX

S. no.	Topic	Page no.
1.	Clustering	8
2.	Dataset summary	8
3.	Decision tree	9
4.	Variable and Properties	9

# Chapter 1 - Introduction

‘BIKE SHARING’ dataset, which we have used in our project for running various models & clustering techniques is basically derived from the UCI Machine Learning repository.

Source: <https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset>

Original Source: <http://capitalbikeshare.com/system-data>

Weather Information: <http://www.freemeteo.com>

Holiday Schedule: <http://dchr.dc.gov/page/holiday-schedule>

This dataset contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information.

This dataset also contains details of various bikes and their rental information. In the original dataset as obtained from the UCI repository was basically .name format in space separated format, from which we applied some normalisations & obtained .csv file with comma separated values. Using this dataset, we can try various things like predicting the name of biker by using their rental information as input features & we can also apply clustering techniques to cluster the various countries on the basis of varied features.

## 1. Attribute Information:

Attributes are used to represent an objects properties. Here we discussed about all the attributes which are essential for a Bike Sharing.

Name, day, season, year , month, hour, holiday, weekday, working day, causal user, registered, count, temperature, windspeed ,weather.

1. instant: record index
2. dteday : date
3. season : season (1:springer, 2:summer, 3:fall, 4:winter)
4. yr : year (0: 2011, 1:2012)
5. mnth : month ( 1 to 12)
6. hr : hour (0 to 23)
7. holiday : weather day is holiday or not (extracted from [Web Link])
8. weekday : day of the week
9. workingday : if day is neither weekend nor holiday is 1, otherwise is 0.
10. weathersit :
  - a. - 1: Clear, Few clouds, Partly cloudy, Partly cloudy

- b. - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  - c. - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  - d. - 4: Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog
11. temp : Normalized temperature in Celsius. The values are derived via  $(t - t_{\min}) / (t_{\max} - t_{\min})$ ,  $t_{\min} = -8$ ,  $t_{\max} = +39$  (only in hourly scale)
  12. atemp: Normalized feeling temperature in Celsius. The values are derived via  $(t - t_{\min}) / (t_{\max} - t_{\min})$ ,  $t_{\min} = -16$ ,  $t_{\max} = +50$  (only in hourly scale)
  13. hum: Normalized humidity. The values are divided to 100 (max)
  14. wind speed: Normalized wind speed. The values are divided to 67 (max)
  15. casual: count of casual users
  16. registered: count of registered users
  17. cnt: count of total rental bikes including both casual and registered.

## Chapter 2 - Technology Used

In our Data Mining Procedures, we have used R. R is a programming language and free software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing. Basically, R is influenced from LISP & scheme.

We installed the complete R Package on our workspace using one of the CRAN Mirrors(Comprehensive R Archive Network). For the GUI Support, we made use of Rattle Library.

### 1. Decision Tree

- a. Decision tree is the most powerful and popular tool for classification and prediction.
- b. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.
- c. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

### 2. Construction of Decision Tree

- a. A tree can be “learned” by splitting the source set into subsets based on an attribute value test.
- b. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions.
- c. The construction of decision tree classifier does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery.
- d. Decision trees can handle high dimensional data. In general decision tree classifier has good accuracy.

### 3. Advantages

- a. Decision Trees are simple enough to understand, interpret its outcome and visualize the results. Able to handle both numeric as well as categorical data and also multi-output problems.
- b. The White box model is followed up. If some situation is observable in the model, then its explanation is easily explained using the logic of Boolean Algebra.

### 4. Disadvantages

- a. Sometimes complex trees are created which are not able to generalize the data well. Decision Trees are prone to Over-fitting.

- b. Decision trees are usually very unstable and even small modifications in the data might lead to an entirely different tree being generated. For the cases, where some classes dominate creation of biased Decision Tree takes place.

## **5. Explanation:**

In our dataset, for each attributes we have some pre assigned values as previously shown in attribute information.

It also gives information of renting bikes in various months under various physical conditions eg. weather, etc.

Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return back has become automatic.

Through these systems, user is able to easily rent a bike from a particular position and return back at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousands bicycles.

Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research.

Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems.

This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of important events in the city could be detected via monitoring these data.

## **Chapter 3 - Result And Applications**

Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return back at another position.

It also gives information of renting bikes in various months under various physical conditions eg. weather, etc.

Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

Through these systems, user is able to easily rent a bike from a particular position and return back at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousands bicycles.

This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of important events in the city could be detected via monitoring these data.

## Chapter 4 - Screenshots

```
Cluster sizes:

[1] "76 176 124 135"

Data means:

      yr      mnth    holiday    weekday workingday weathersit      temp
0.49510763 0.49439601 0.03131115 0.50880626 0.68493151 0.21037182 0.54227350
      atemp      hum  windspeed      casual registered      cnt
0.50881294 0.56205474 0.35257901 0.23958821 0.51996458 0.50652466

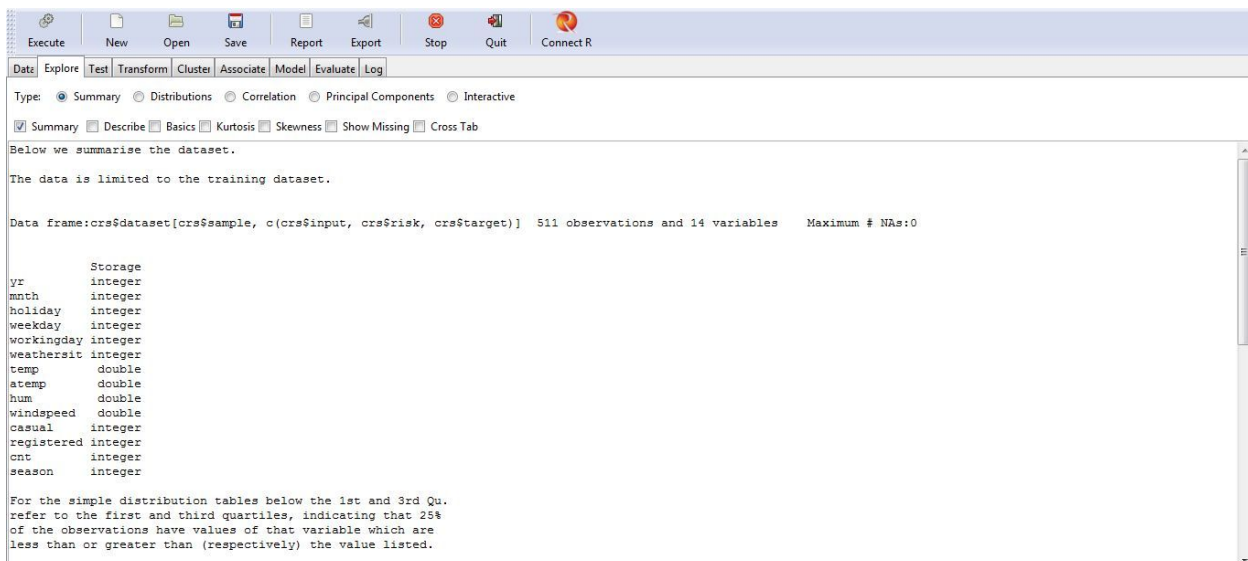
Cluster centers:

      yr      mnth    holiday    weekday workingday weathersit      temp
1 0.0000000 0.1112440 0.01315789 0.5394737 0.6578947 0.2565789 0.2703507
2 1.0000000 0.4943182 0.00000000 0.5208333 1.0000000 0.1960227 0.5703400
3 0.0000000 0.6436950 0.00000000 0.5147849 1.0000000 0.2338710 0.6384074
4 0.5703704 0.5730640 0.11111111 0.4703704 0.0000000 0.1814815 0.5704648
      atemp      hum  windspeed      casual registered      cnt
1 0.2540238 0.5341035 0.4090612 0.0663300 0.2004094 0.1850295
2 0.5330813 0.5336518 0.3451042 0.2166493 0.7217565 0.6576502
3 0.6006156 0.6075466 0.3436674 0.1716289 0.4922184 0.4578626
4 0.5362886 0.5730342 0.3387121 0.4294536 0.4622708 0.5351886

Within cluster sum of squares:

[1] 42.91810 77.58944 44.20093 137.65005
```

Figure 1. - Clustering



Execute New Open Save Report Export Stop Quit Connect R

Data Explore Test Transform Cluster Associate Model Evaluate Log

Type: ☒ Summary ☐ Distributions ☐ Correlation ☐ Principal Components ☐ Interactive

☒ Summary ☐ Describe ☐ Basics ☐ Kurtosis ☐ Skewness ☐ Show Missing ☐ Cross Tab

Below we summarise the dataset.

The data is limited to the training dataset.

Data frame: crrs[crrs\$sample, c(crrs\$input, crrs\$risk, crrs\$target)] 511 observations and 14 variables Maximum # NAs: 0

Variable	Storage
yr	integer
mnth	integer
holiday	integer
weekday	integer
workingday	integer
weathersit	integer
temp	double
atemp	double
hum	double
windspeed	double
casual	integer
registered	integer
cnt	integer
season	integer

For the simple distribution tables below the 1st and 3rd Qu. refer to the first and third quartiles, indicating that 25% of the observations have values of that variable which are less than or greater than (respectively) the value listed.

Figure 2. - Dataset summary



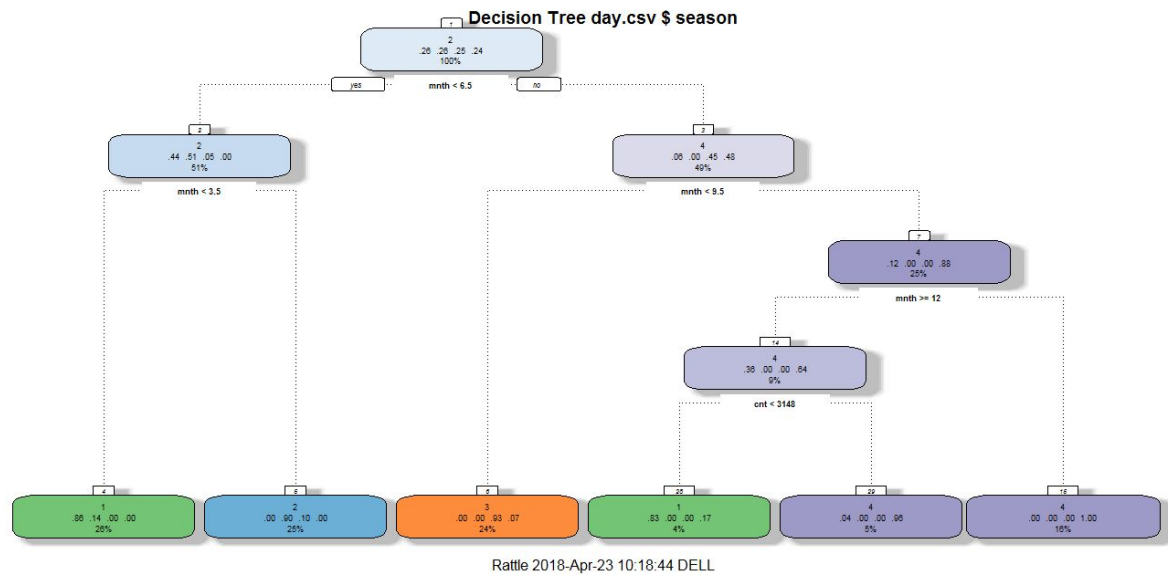


Figure 3. - Decision tree

Execute New Open Save Report Export Stop Quit Connect R

Date Explore Test Transform Cluster Associate Model Evaluate Log

Source: ☒ File ☐ ARFF ☐ ODBC ☐ R Dataset ☐ RData File ☐ Library ☐ Corpus ☐ Script

Filename: day.csv Separator: Decimal: Header

☒ Partition 70/15/15 Seed: 42 View Edit

☒ Input ☐ Ignore Weight Calculator: Target Data Type: ☒ Auto ☐ Categorical ☐ Numeric ☐ Survival

No. Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1 instant	Ident	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 731
2 dteday	Ident	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 731
3 season	Numeric	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 4
4 yr	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 2
5 mnth	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 12
6 holiday	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 2
7 weekday	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 7
8 workingday	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 2
9 weathersit	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 3
10 temp	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 499
11 atemp	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 690
12 hum	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 595
13 windspeed	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 650
14 casual	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 606
15 registered	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 679
16 cnt	Numeric	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Unique: 696

Figure 4. - Variable and properties