# Producing Standard Rules for Smart Real Estate Property Buying Decisions based on Web Scraping Technology and Machine Learning Techniques

Haris Ahmed*[1], Tahseen Ahmed Jilani[2], Waleej Haider[3], Syed Noman Hasany[4]
Mohammad Asad Abbasi[5], Ahsan Masroor[6]
Department of Computer Science, Sir Syed University of Engineering and Technology, Karachi, Pakistan[1, 3, 4, 5, 6]
Department of Computer Science, University of Karachi, Karachi, Pakistan[2]
School of Computer Science, University of Nottingham, Nottingham, UK[2]

***Abstract*—Purchasing of real estate property is a stressful and time-consuming activity, regardless of the individual in question is a buyer or seller. The act is also a major financial decision which can lead to numerous consequences if taken hastily. Therefore, it is encouraged that a person should properly invest their time and money in research relating to price demands, property type and location, etc. It can be a difficult task to assess what real estate property can be considered as the best property to buy. The key idea of the current research study is to create a set of standard rules, which should be embraced to make a smart decision of buying real estate property, based on web scraping technology and machine learning techniques.**

***Keywords*—Web scraping technology; HtmlAgilityPack; machine learning; C4.5 decision tree; Weka-J48***

## I. Introduction

Any decision in relation to a property purchase or sales is a vital decision. To say that it is difficult to make up one's mind in that circumstance is an understatement [1]. However, that is not to say as it is impossible to do so as there are technological means available to the modern man that allow them to make the best decision. One such route is to take the assistance of web scraping technology. This form of tech allows the user to find various online real state property advertisements from different web sources [2]. Therefore, the individual will have a much better idea of what sort of decision they should be making in terms of selling or buying real estate. Furthermore, with the help of machine learning techniques such as decision tree C4.5 [3], in combination with the prior mentioned option, one can easily make a superior decision.

## II. Web Scraping using HTML Agility Pack

The term "Web Scraping" also referred to as the "screen scraping or web data extraction technique" is a program for mining huge volume of data from an internet source, removing the information and saving it to a local file in a computer or databank and it saves the table in a spreadsheet format [4].

The data exhibited on numerous internet sources can only be observed through a web browser. Therefore, the sole possibility is to physically copy-paste the information. This is a very monotonous task that can take a lot of time, even days to complete. In addition to this, web crawling is a procedure that mechanizes this process. As a result, Web Scraping software does not need to manually copy data from a source but can perform the same task quickly.

### A. HTML Agility-Pack

This is a responsive "HTML parser" inscribed in C# that builds a read/write "DOM" and supports basic "XPATH" or "XSLT" [5]. It is a ".NET code library" that permits you to analyse "out of the web" HTML archives. For improved understanding, "HTML Agility pack" is used to contrivance scraping of several web pages present on the internet [6].

- HTML Parsing

HTML parsing is fundamental as taking in HTML code and mining applicable data like the title of the page, subsections in the page, relations, bold text etc.

- Document Object Model

The "Document Object Model" is a software design "API" for "HTML" and "XML" documents. It outlines the rational construction of documents and the method by which a document is retrieved and deployed [7].

### B. HTML Agility Pack Installation Steps

*1)* First, install the "NuGet package".

*2)* Below the segment "Package Manager" copy the installed code. Such as, if there is a statement of "PM> Install-Package HtmlAgilityPack - Version 1.5.1," then the text following the "PM>" shall be copied by the user.

*3)* Afterwards, go to the "Visual Studio Application" and click on "Tools menu" in the menu bar.

*4)* Using the drop-down menu, go to the library "manager Package Manager Console."

*5)* Starting from the bottom, "Application," the "Package Manager Console" opened and the cursor blinking.

*6)* The copied code should be pasted from the internet site through the "help of step 2" using hotkeys "Ctrl and V."

*7)* Press enter and the application will install automatically.

*Corresponding Author

### C. Steps To Load DOM Using HTML Agility Pack

*1)* Add a DLL reference by going into the "Visual Studio Application" and press on the "Solution Explorer" positioned in the sidebar.

*2)* Right-click on and then click on "Add Reference," in the context menu.

*3)* From the "Reference Manager window," click on the "browser button" and move to "HAP dll" to select it.

*4)* Press Ok and go back to the code area of the "Visual Studio application" and insert desired code.

*5)* Inside the Main-Function, write the following code.

HTML Agility Pack will be used to load the HTML Document

```
HtmlWeb web = new HtmlWeb();
HtmlAgilityPack.HtmlDocument doc = new
HtmlAgilityPack.HtmlDocument();
doc = web.Load("http://technologyCrowds.com");
GetMetaInformation(doc,"description");
```

"GetMetaInformation" method definition.

```
static void GetMetaInformation
(HtmlAgilityPack.HtmlDocument htmldoc, string
value)
{
 HtmlNode tcNode =
htmldoc.DocumentNode.SelectSingleNode("//meta
[@name='" + value + "']");
 string fulldescription = string.Empty;
if(tcNode != null)
{
 HtmlAttribute desc;
 desc = tcNode.Attributes["content"];
 Console.ForegroundColor = ConsoleColor.Red;
 Console.Write(desc.Value);
 Console.ReadLine();
 }
}
```

*6)* For the main function, the user should click on "Start Button" after saving the code and place cursor on the line "doc=web.load" ("https://technologycrowds.com"); and click on "DocumentNode," then "InnerHtml."

*7)* Click on the "search icon" and a new window will pop up. The new window will have all the "DOM" contents which are "HTML" content.

## III. Research Methodology

In the first step we have to briefly list the "URL addresses" of the best online real estate ad web sources, then pass all the URLs in "HtmlAgilityPack" to extract the real estate ad data (e.g. property positions, prices and publication date of the ads, etc.) from numerous web sources. In the next step, with the help of linear regression, we will find the average future growth rate of the prices of each real estate property. In the end, with changes in the current average property prices and the estimated average future growth rates, we create a set of standard rules for making decisions about buying a real estate property. Fig. 1 shows the steps of the research methodology.

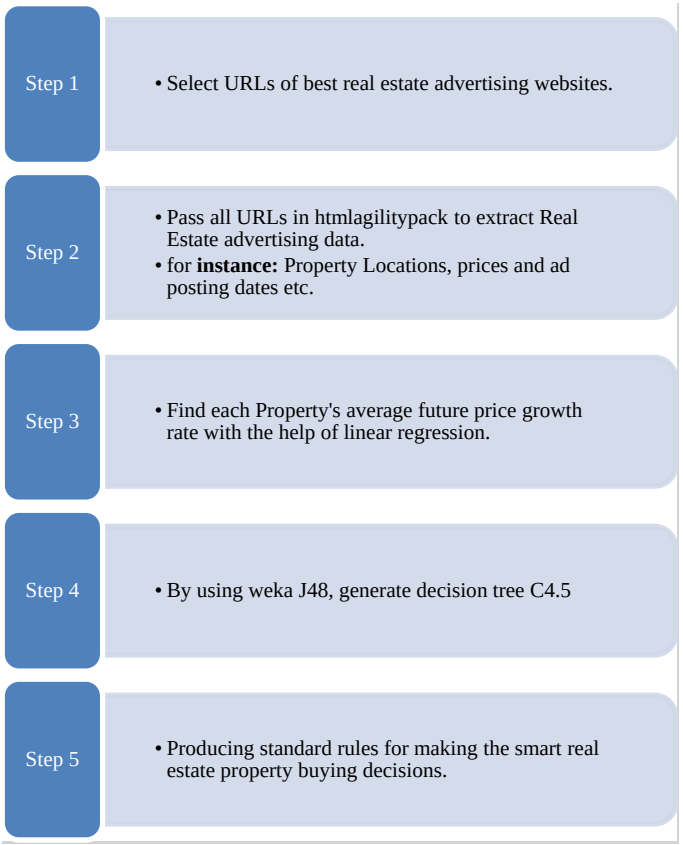


Fig. 1. Steps for Executing the Methodology.

## IV. Extraction Real Estate Advertisement Data from Various Web Sources

The particular research has used the web-scraping technology i.e. HTML Agility Pack, which uses the Pakistan Online Real Estate websites and their advertisements to bring the desired results. The chosen results are brought by the help of web-scraping technology and Table I shows the average prices of the most popular housing areas in different periods of time.

The future price growth rate of any real estate property has become a very significant factor in order to make real estate property buying decisions. The average prices were seen in the below table i.e. Table I and by the help of those average prices of different time intervals, we can use the linear regression technique to assess the average growth rate of future real estate prices.

TABLE. I. AVERAGE PRICES OF POPULAR LOCATION ON DIFFERENT INTERVALS OF TIME

| Sr. No. | Popular Locations for Houses | Avg. Prices From Year 2014 To 2015 (Millions In PKR) | Avg. Prices From the Year 2016 To 2017 (Millions In PKR) | Avg. Prices From Year 2018 To 2019 (Millions In PKR) |
|---|---|---|---|---|
| 1 | Gulshan-e-Iqbal Karachi | 56 | 65 | 74 |
| 2 | Gulistan-e-jauhar Karachi | 41 | 50 | 58 |
| 3 | Shahra-e-Faisal Karachi | 40 | 51 | 59 |
| 4 | Gulberg Lahore | 55 | 60 | 65 |
| 5 | Cantt Lahore | 38 | 40 | 43 |
| 6 | Gulberg Islamabad | 42 | 42 | 44 |
| 7 | Kashmir Highway Islamabad | 100 | 106 | 109 |
| 8 | Lalazar Rawalpindi | 7 | 10 | 14 |
| 9 | Gulshan Abad Rawalpindi | 9 | 9 | 10 |
| 10 | Saddar Rawalpindi | 21 | 20 | 21 |
| 11 | North Karachi Karachi | 9 | 10 | 15 |
| 12 | North Nazimabad Karachi | 50 | 54 | 55 |
| 13 | Malir Karachi | 12 | 14 | 15 |
| 14 | Cantt Karachi | 58 | 59 | 60 |
| 15 | Mehmoodabad Karachi | 2 | 4 | 7 |
| 16 | Ghauri Town Islamabad | 12 | 13 | 14 |
| 17 | Kuri Road Islamabad | 20 | 21 | 20 |
| 18 | Bharakahu Islamabad | 55 | 60 | 70 |
| 19 | Simly Dam Road Islamabad | 89 | 93 | 98 |
| 20 | GT Road Islamabad | 7 | 7 | 7 |
| 21 | Harbanspura Lahore | 8 | 8 | 9 |
| 22 | Ferozepur Road Lahore | 2 | 3 | 5 |
| 23 | Taj Pura Lahore | 2 | 3 | 6 |
| 24 | Walton Road Lahore | 5 | 8 | 11 |
| 25 | Gulshan-e-Ravi Lahore | 25 | 32 | 40 |
| 26 | Misryal Road Rawalpindi | 9 | 10 | 10 |
| 27 | Shakrial Rawalpindi | 2 | 3 | 5 |
| 28 | Sadiqabad Rawalpindi | 6 | 8 | 10 |

## V. SIMPLE LINEAR REGRESSION

Simple linear regression establishes the connection between "target variable" and "input variables" by fitting a line, called "regression line" [8]. Generally, the linear equation

$$y = m * x + b \quad (1)$$

The above equation is used to represent the line. Within the equation, "y" acts as the dependent variable, whereas "x" is the independent. "m" depicts the "slope", and "b" is the "intercept point".

Machine learning requires the following iteration of the same equation.

$$y(x) = w0 + w1 * x \quad (2)$$

Where "w" denotes the parameters, "x" acts as the input, and "y" is the target variable. Changing values of w0 and w1 will give us different lines, as seen in Fig. 2.

Based on "Linear Regression Analysis" Table II offered the estimated average future property values in the different lengths of time and price growth rate percentage.

As a portion of pre-processing the constant assessed real estate records shown in Table II is renewed to definite form by estimated width of the preferred intervals, as shown in Table III.

Table IV visibly demonstrates the projected real estate property data set that is converted into the categorical form.

Next, the categorical data is given as input to "Decision tree C4.5" (Weka-J4.8)

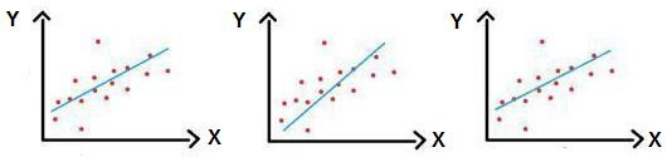


Fig. 2. Linear Regression Lines.

TABLE. II. ESTIMATED AVERAGE FUTURE PROPERTY PRICE GROWTH RATE

| Sr. No. | Popular Locations For Houses | Estimated Avg. Future Prices From Year 2020 To 2021 (Millions In PKR) | Estimated Avg. Future Prices From Year 2022 To 2023 (Millions In PKR) | Estimated Avg. Future Prices From Year 2024 To 2025 (Millions In PKR) | Estimated Average Future Price Growth Rate Percentage |
|---|---|---|---|---|---|
| 1 | Gulshan-e-Iqbal Karachi | 81 | 90 | 98 | 10% |
| 2 | Gulistan-e-jauhar Karachi | 65 | 73 | 81 | 12% |
| 3 | Shahra-e-Faisal Karachi | 67 | 76 | 85 | 13% |
| 4 | Gulberg Lahore | 69 | 74 | 78 | 6% |
| 5 | Cantt Lahore | 45 | 47 | 49 | 4% |
| 6 | Gulberg Islamabad | 44 | 45 | 46 | 2% |
| 7 | Kashmir Highway Islamabad | 113 | 117 | 121 | 3% |
| 8 | Lalazar Rawalpindi | 17 | 20 | 23 | 16% |
| 9 | Gulshan Abad Rawalpindi | 10 | 11 | 11 | 5% |
| 10 | Saddar Rawalpindi | 21 | 21 | 21 | 0% |
| 11 | North Karachi Karachi | 17 | 20 | 22 | 14% |
| 12 | North Nazimabad Karachi | 58 | 60 | 62 | 3% |
| 13 | Malir Karachi | 16 | 18 | 19 | 9% |
| 14 | Cantt Karachi | 61 | 62 | 63 | 2% |
| 15 | Mehmoodabad Karachi | 9 | 11 | 13 | 20% |
| 16 | Ghauri Town Islamabad | 15 | 16 | 17 | 6% |
| 17 | Kuri Road Islamabad | 20 | 20 | 20 | 0% |
| 18 | Bharakahu Islamabad | 75 | 82 | 89 | 9% |
| 19 | Simly Dam Road Islamabad | 102 | 106 | 110 | 4% |
| 20 | GT Road Islamabad | 7 | 7 | 7 | 0% |
| 21 | Harbanspura Lahore | 9 | 10 | 10 | 6% |
| 22 | Ferozepur Road Lahore | 6 | 7 | 9 | 23% |
| 23 | Taj Pura Lahore | 7 | 9 | 11 | 25% |
| 24 | Walton Road Lahore | 13 | 16 | 19 | 21% |
| 25 | Gulshan-e-Ravi Lahore | 46 | 53 | 60 | 14% |
| 26 | Misryal Road Rawalpindi | 11 | 11 | 11 | 0% |
| 27 | Shakrial Rawalpindi | 6 | 7 | 9 | 23% |
| 28 | Sadiqabad Rawalpindi | 12 | 13 | 15 | 12% |

TABLE. III. CATEGORICAL-PARTITIONING OF ESTIMATED REAL ESTATE PROPERTY DATA SET

| Sr. No | Popular Locations For Houses | Partitioned Data | |
|---|---|---|---|
| | | Avg. Current Price Rate Year 2019 (Millions In PKR) | Estimated Avg. Future Price Growth Rate Percentage |
| 1 | Gulshan-e-Iqbal Karachi | {low, medium, high}<br>{ <55,55-60,>60} | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 2 | Gulistan-e-jauhar Karachi | {low, medium, high}<br>{ <55,55-60,>60 } | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 3 | Shahra-e-Faisal Karachi | {low, medium, high}<br>{ <55,55-60,>60 } | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 4 | Gulberg Lahore | {low, medium, high}<br>{ <70,70-75,>75 } | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 5 | Cantt Lahore | {low, medium, high}<br>{ <40,40-45,>45} | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 6 | Gulberg Islamabad | {low, medium, high}<br>{ <45,45-50,>50 } | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 7 | Kashmir Highway Islamabad | {low, medium, high}<br>{ <110,110-120,>120} | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 8 | Lalazar Rawalpindi | {low, medium, high}<br>{<10,10-15,>15 } | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 9 | Gulshan Abad Rawalpindi | {low, medium, high}<br>{ <5,5-10,>10 } | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 10 | Saddar Rawalpindi | {low, medium, high}<br>{ <20,20-25,>25 } | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 11 | North Karachi Karachi | {low, medium, high}<br>{ <5,5-10,10> } | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 12 | North Nazimabad Karachi | {low, medium, high}<br>{ <45,45-50,>50 } | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 13 | Malir Karachi | {low, medium, high}<br>{ <15,15-20,>20 } | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 14 | Cantt Karachi | {low, medium, high}<br>{ <55,55-60,>60 } | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 15 | Mehmoodabad Karachi | {low, medium, high}<br>{ <4,4-6>6 } | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 16 | Ghauri Town Islamabad | {low, medium, high}<br>{ <5,5-10,>10 } | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 17 | Kuri Road Islamabad | {low, medium, high}<br>{ <25,25-30,>30 } | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 18 | Bharakahu Islamabad | {low, medium, high}<br>{ <65,65-70,>70} | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 19 | Simly Dam Road Islamabad | {low, medium, high}<br>{ <70,70-80,>80 } | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 20 | GT Road Islamabad | {low, medium, high}<br>{ <5,5-10,>10 } | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 21 | Harbanspura Lahore | {low, medium, high}<br>{ <10,10-15,>15 } | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 22 | Ferozepur Road Lahore | {low, medium, high}<br>{ <6,6-8,>8 } | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 23 | Taj Pura Lahore | {low, medium, high}<br>{ <3,3-5,>5 } | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 24 | Walton Road Lahore | {low, medium, high}<br>{ <15,15-20,>20 } | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 25 | Gulshan-e-Ravi Lahore | {low, medium, high}<br>{ <25,25-30,>30 } | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 26 | Misryal Road Rawalpindi | {low, medium, high}<br>{ <6,6-8,>8 } | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 27 | Shakrial Rawalpindi | {low, medium, high}<br>{ <6,6-8,>8 } | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |
| 28 | Sadiqabad Rawalpindi | {low, medium, high}<br>{ <15,15-20,>20} | {weak, moderate, strong}<br>{ <5%,5-10%,>10% } |

TABLE. IV. CATEGORICAL REAL ESTATE PROPERTY DATA SET

| Sr. No | Popular Locations For Houses | Avg. Current Price Rate | Estimated Avg. Future Price Growth Rate |
|---|---|---|---|
| 1 | Gulshan-e-Iqbal Karachi | High | Moderate |
| 2 | Gulistan-e-jauhar Karachi | Medium | Strong |
| 3 | Shahra-e-Faisal Karachi | Medium | Strong |
| 4 | Gulberg Lahore | Low | Moderate |
| 5 | Cantt Lahore | Medium | Weak |
| 6 | Gulberg Islamabad | Low | Weak |
| 7 | Kashmir Highway Islamabad | Low | Weak |
| 8 | Lalazar Rawalpindi | Medium | Strong |
| 9 | Gulshan Abad Rawalpindi | Medium | Moderate |
| 10 | Saddar Rawalpindi | Medium | Weak |
| 11 | North Karachi Karachi | High | Strong |
| 12 | North Nazimabad Karachi | High | Weak |
| 13 | Malir Karachi | Medium | Moderate |
| 14 | Cantt Karachi | Medium | Weak |
| 15 | Mehmoodabad Karachi | High | Strong |
| 16 | Ghauri Town Islamabad | High | Moderate |
| 17 | Kuri Road Islamabad | Low | Weak |
| 18 | Bharakahu Islamabad | Medium | Moderate |
| 19 | Simly Dam Road Islamabad | High | Weak |
| 20 | GT Road Islamabad | Medium | Weak |
| 21 | Harbanspura Lahore | Low | Moderate |
| 22 | Ferozepur Road Lahore | Low | Strong |
| 23 | Taj Pura Lahore | High | Strong |
| 24 | Walton Road Lahore | Low | Strong |
| 25 | Gulshan-e-Ravi Lahore | High | Strong |
| 26 | Misryal Road Rawalpindi | High | Weak |
| 27 | Shakrial Rawalpindi | Low | Strong |
| 28 | Sadiqabad Rawalpindi | Low | Strong |

## VI. DECISION TREE C4.5

Decision tree refers to a "supervised classification method" that is a structure in which the non-terminal nodes indicate the test of one or more features, and the terminal nodes indicate the result of the decision [9]. It has been apprehended from the studies that the basic algorithm for determining the tree ID3 derivation has been enhanced by the C4.5 algorithm [10]. The unique C4.5 version called J4.8 has a WEKA classification package [11]. In C4.5, the information gain ratio and its measurements are used as a splitting principle, respectively [12]. The steps of this algorithm are given as follows:

Step 1: The set 't' is a set of class labels for tuple training. If an output test is selected, the sample 't' training set must be split into subsets {T1, T2...Tn}. So, the entropy of the set-T can be calculated (in bits).

$$info(T) = -\sum_{i=1}^{k}((freq(C_i,T)/|T|) \times \log_2(freq(C_i,T)/|T|)) \quad (3)$$

Step 2: Divide the training sample by the value of the specified attribute, by which the value of property T will be:

$$infox(T) = -\sum_{i=1}^{n}((|T_i|/|T|) \times info(T_i)) \quad (4)$$

Step 3: Afterwards, the difference between basic information requirements and new information is referred by the information gain. The equation (3) and equation (4) can provide a gain standard:

$$Gain(X) = info(T) - infox(T) \quad (5)$$

Step 4: When building a dense decision tree, the quality of the gain is beneficial, but the test has significant disadvantages because many outputs have large deviations. Therefore, it has to be determined by standardization:

$$Split - info(X) = -\sum_{i=1}^{n}((|T_i|/|T|)\log_2(|T_i|/|T|)) \quad (6)$$

The new gain standard is represented as:

$$Gain - ratio(X) = gain(X)/split - info(x) \quad (7)$$

The Real Estate training dataset (see Table V) is provided as input to "WekaJ48".

TABLE. V. REAL ESTATE TRAINING DATA SET

| Sr. No | Popular Locations For Houses | Avg. Current Price Rate | Estimated Avg. Future Price Growth Rate | Class |
|---|---|---|---|---|
| 1 | Gulshan-e-Iqbal Karachi | High | Moderate | No |
| 2 | Gulistan-e-jauhar Karachi | Medium | Strong | Yes |
| 3 | Shahra-e-Faisal Karachi | Medium | Strong | Yes |
| 4 | Gulberg Lahore | Low | Moderate | Yes |
| 5 | Cantt Lahore | Medium | Weak | No |
| 6 | Gulberg Islamabad | Low | Weak | No |
| 7 | Kashmir Highway Islamabad | Low | Weak | No |
| 8 | Lalazar Rawalpindi | Medium | Strong | Yes |
| 9 | Gulshan Abad Rawalpindi | Medium | Moderate | No |
| 10 | Saddar Rawalpindi | Medium | Weak | No |
| 11 | North Karachi Karachi | High | Strong | No |
| 12 | North Nazimabad Karachi | High | Weak | No |
| 13 | Malir Karachi | Medium | Moderate | No |
| 14 | Cantt Karachi | Medium | Weak | No |
| 15 | Mehmoodabad Karachi | High | Strong | No |
| 16 | Ghauri Town Islamabad | High | Moderate | No |
| 17 | Kuri Road Islamabad | Low | Weak | No |
| 18 | Bharakahu Islamabad | Medium | Moderate | No |
| 19 | Simly Dam Road Islamabad | High | Weak | No |
| 20 | GT Road Islamabad | Medium | Weak | No |
| 21 | Harbanspura Lahore | Low | Moderate | Yes |
| 22 | Ferozepur Road Lahore | Low | Strong | Yes |
| 23 | Taj Pura Lahore | High | Strong | No |
| 24 | Walton Road Lahore | Low | Strong | Yes |
| 25 | Gulshan-e-Ravi Lahore | High | Strong | No |
| 26 | Misryal Road Rawalpindi | High | Weak | No |
| 27 | Shakrial Rawalpindi | Low | Strong | Yes |
| 28 | Sadiqabad Rawalpindi | Low | Strong | Yes |

## VII. STAGES TO CREATE "C4.5 DECISION TREE" IN "WEKA J48"

*1)* Generate datasets in "MS Excel," "MS Access" and save in "CSV" format.

*2)* Start the "weka Explorer."

*3)* Open ".CSV" file and change format to "ARFF."

## VIII. RULE PRODUCTION USING DECISION-TREE [14,15] FOR MAKING SMART REAL ESTATE PROPERTY BUYING DECISIONS

The corresponding rules are:

R1: IF (Estimated Avg. Future Price Growth Rate=Weak) Then Purchase = No

R2: IF (Estimated Avg. Future Price Growth Rate=Strong) AND (Avg. Current Price Rate=Low) THEN Purchase = Yes

R3: IF (Estimated Avg. Future Price Growth Rate=Strong) AND (Avg. Current Price Rate=Medium) THEN Purchase = Yes

R4: IF (Estimated Avg. Future Price Growth Rate=Strong) AND (Avg. Current Price Rate=High) THEN Purchase = No

R5: IF (Estimated Avg. Future Price Growth Rate=Moderate) AND (Avg. Current Price Rate=Low) THEN Purchase = Yes

R6: IF (Estimated Avg. Future Price Growth Rate=Moderate) AND (Avg. Current Price Rate=Medium) THEN Purchase = No

R7: IF (Estimated Avg. Future Price Growth Rate=Moderate) AND (Avg. Current Price Rate=High) THEN Purchase = No

*1)* Click "classify tab," then select "J48."

*2)* Select any suitable test possibility.

*3)* Click "Start".

Click on "Visualize Tree" option to view the graphical representation of the tree from the pop-up menu. Fig. 3 depicts the graphical form of Weka J48 generated tree [13].

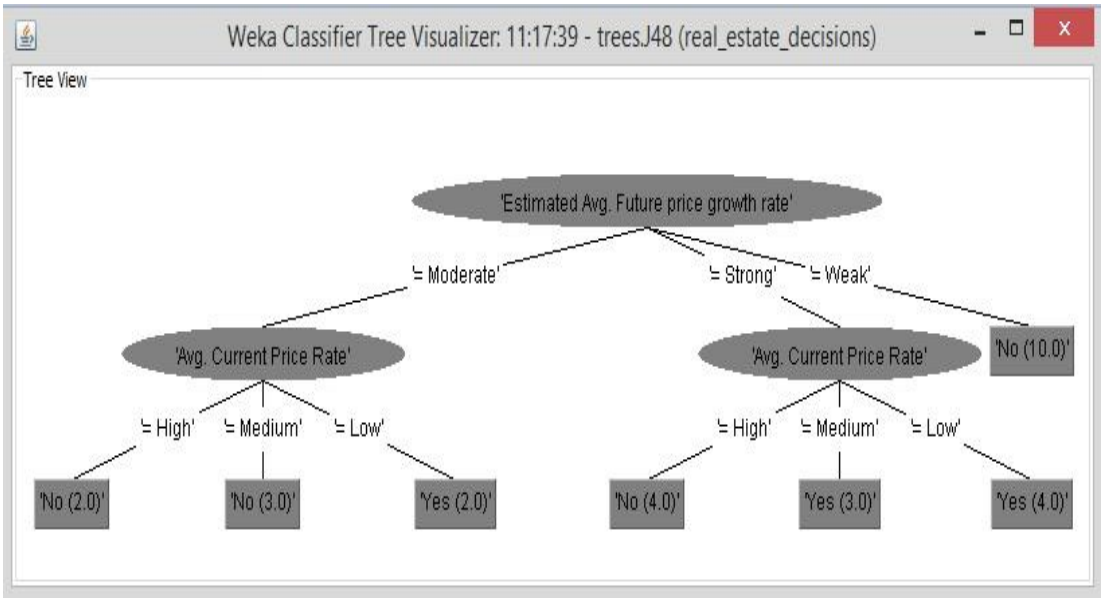


Fig. 3. Real Estate Property Buying Decision Tree C4.5 using Weka J48.

These rules are classified into two classes "YES" and "NO". The following study discloses only one of the decision-rule for each class.

### A. *"NO" Class Rule*

R1: IF (Estimated Avg. Future Price Growth Rate=Weak) Then Purchase = No

It specifies that when the Estimated Avg. Future Price Growth Rate=Weak then purchasing of real estate property is not advisable. Furthermore, 10 training examples support the rule.

### B. *"YES" Class Rule*

R2: IF (Estimated Avg. Future Price Growth Rate=Strong) AND (Avg. Current Price Rate=Low) THEN Purchase = Yes

It specifies that when the Estimated Avg. Future Price Growth Rate=Strong and Avg. Current Price Rate=Low then purchasing of real estate property is beneficial. Furthermore, 4 training examples support the rule.

## IX. Conclusion

The decision to buy real estate is a substantial financial decision. Buyers should spend a lot of time choosing the best property to buy from all available options. This research concludes that there are no existing standard rules for making smart real estate purchase decisions. However, we propose a method that can generate standard rules for selecting the best real estate property to buy through web scraping technology and machine learning algorithms. This research will save buyers' time and provide a complete guide to make smart real estate buying decisions.

## Acknowledgment

Authors would like to thank Center of Innovations in Computer Science (CICS), Sir Syed University of Engineering and Technology for providing resources and support to perform experiments. This work is also supported by the University of Nottingham, UK.

## References

[1] Brueggeman, William B., and Jeffrey D. Fisher. Real estate finance and investments. New York, NY: McGraw-Hill Irwin, 2011.

[2] Vargiu, Eloisa, and Mirko Urru. "Exploiting web scraping in a collaborative filtering-based approach to web advertising." Artif. Intell. Research 2.1 (2013): 44-54.

[3] Hssina, Badr, et al. "A comparative study of decision tree ID3 and C4. 5." International Journal of Advanced Computer Science and Applications 4.2 (2014): 13-19.

[4] Wijaya, James. "Ekstraksi Teks Pada Halaman Website Renungan Rohani Menggunakan HTML Agility Pack." (2019).

[5] https://html-agility-pack.net/

[6] Uzun, Erdinç, et al. "Evaluation of Hap, AngleSharp and HtmlDocument in web content extraction." International Scientific Conference'2017 (UNITECH'17). 2017.

[7] Álvarez-Sabucedo, L. M., Luis E. Anido-Rifón, and Juan M. Santos-Gago. "Reusing web contents: a DOM approach." Software: Practice and Experience 39.3 (2009): 299-314.

[8] Bangdiwala, Shrikant I. "Regression: simple linear." International journal of injury control and safety promotion 25.1 (2018): 113-115.

[9] Siahaan, Hasudungan, et al. "Application of Classification Method C4. 5 on Selection of Exemplary Teachers." Journal of Physics: Conference Series. Vol. 1235. No. 1. IOP Publishing, 2019.

[10] Sathyadevan, Shiju, and Remya R. Nair. "Comparative analysis of decision tree algorithms: ID3, C4. 5 and random forest." Computational intelligence in data mining-volume 1. Springer, New Delhi, 2015. 549-562.

[11] Bhargava, Neeraj, et al. "Decision tree analysis on j48 algorithm for data mining." Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering 3.6 (2013).

[12] Quinlan, J. Ross.C4. 5: programs for machine learning. Elsevier, 2014

[13] Drazin, Sam, and Matt Montag. "Decision tree analysis using weka." Machine Learning-Project II, University of Miami (2012): 1-3.

[14] Jain, Rajni. "Rule generation using decision trees." *IASRI* (2012).

[15] Al-Radaideh, Qasem A., Emad M. Al-Shawakfa, and Mustafa I. Al-Najjar. "Mining student data using decision trees." International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan. 2006.