Group No - 32

# MAJOR PROJECT II

**FINAL REPORT**

**ON**

## Web Scraping And Prediction Of Selling Prices Using BeautifulSoup And Random Forest Regressor

**Submitted By**


| Sandeep Kumar | Rishank Gupta | Mayank Shekhar | Bedabandhu Sahoo |
|---|---|---|---|
| 500062372 | 500062486 | 500063785 | 500063049 |


***Under the guidance of***

**Dr. Rohit Tanwar**


Assistant Professor - Selection Grade
Department of Systemics,
School of Computer Science


UPES
UNIVERSITY WITH A PURPOSE


**Department of Cybernetics,**
**School of Computer Science**
**UNIVERSITY OF PETROLEUM AND ENERGY STUDIES**
**Dehradun-248007**

**May - 2021**

**PROJECT TITLE:** Web Scraping And Prediction Of Selling Prices Using BeautifulSoup And Random Forest Regressor.

## ABSTRACT:

Information Analysis has been an extraordinary assistance in understanding information in a few zones like financial exchange, endeavors, climate, power interest, cost and utilization of items like fuel, power, and so on It outfits relationship with significant information that is essential to make taught decisions.

In this venture we pursue fabricating our own informational collection shaped by scratching sites. Further, the educational list is to be cleaned, to take out the various irregularities that arise in the assortment and is to be imagined. After that further suitable investigation is done on the information like expectation of MRP of vehicles.

***Keywords** - Data Analysis, scraping, data-set, Machine Learning, WebApp.

## CONTENT

## INTRODUCTION:

Whether you're a scientist analysing earthquake data to predict the next "big one", or are in health-care analysing patient wait times to better staff your ER, understanding data is crucial to making better, data informed decisions. However, you need to run over this information in a productive way, so information is gotten in less time so that additional time can be spent on dissecting this gathered information. Although one may have collected data, cleaning it becomes more important to have a much more unambiguous data set.
Data collection is gathering (from relevant sources), the various measures and information related to certain variables in question present in a system. Information can be gathered by different methods like meetings to generate new ideas, interviews, reviews, contextual investigations, web publicizing and so on. In this project, we look at gathering data by scraping a website.

Web scraping is used to collect large information from websites. It is a computerized strategy used to extricate a lot of information from sites. The data on the websites are unstructured. Web scratching helps gather these unstructured information and store it in an organized structure. There are various approaches to scratch sites like online Services, APIs or composing your own code **[1]**.

Data Cleaning is the process of identifying and removing errors in the data. While gathering and joining information from different sources into an information stockroom, guaranteeing high information quality and consistency turns into a huge, frequently costly and continually testing task. Without spotless and right information the handiness of Data Mining and information warehousing is relieved. This paper examines the issue of information purging and the ID of expected blunders in informational indexes **[2]**. So the point becomes to improve the information quality.

Regular information quality problems(anomalies) incorporate conflicting information shows among sources like various shortenings or equivalent words; information section blunders like spelling botches, conflicting information designs, missing, deficient, obsolete or in any case mistaken characteristic qualities, information duplication, insignificant items or information. Data that is incomplete or inaccurate is known as dirty data **[3]**. The different kinds of irregularities happening in information that must be wiped out. The kind of oddities can be

ordered under a few sorts of it. In view of this arrangement we assess and contrast existing methodologies for information purging and regard to the sorts of inconsistencies took care of and dispensed with by them **[3]**.

Information cleaning offers the principal administrations for information cleaning, for example, characteristic determination, arrangement of tokens, choice of grouping calculation, choice of comparability work, choice of end work and union capacity and so forth. **[3]**. To clean data we make use of Tableau software.
After the cleaning part we utilized irregular woodland regressor and utilized various ascribes accessible to us to infer connection among them and the selling cost of vehicles, to anticipate our qualities and contrast them with our accessible dataset.

## PROBLEM STATEMENT:

To choose a car ,which is also a hefty investment, a lot of time is wasted by every individual to reduce that time we used random forest regressor to help make better and fast choices, the model makes use of multiple attributes of car like facilities, engine, transmission, safety etc to make a differentiated choice.

## OBJECTIVE:

- Gathering data by web scraping in Python.
- Cleaning the gathered data and converting it into a dataset.
- Analysis on the basis of the gathered dataset.
- Splitting the dataset for testing and training.
- Predicting the MRP of cars and comparing them to testing data.
- Creating a WebApp that will show Output.

## OBJECTIVE ACHIEVED:

- Gathered data by web scraping in Python.
- Cleaning the gathered data and converted it into a dataset.
- Analysing on the basis of the gathered dataset.
- Successfully predicted the MRP of cars after applying regression.
- Created a WebApp GUI for Output.

## METHODOLOGY:

We will use a combination of iterative and incremental process models (Agile SDLC model) with focus on process adaptability. This will break the project into small incremental builds. These builds are provided in iterations. Each iteration will last from about one to three weeks.

- Requirement Analysis
- Gathering Data
- Building Data Set
- Cleaning Data Set
- Pre processing of data
- Visualizing results
- applying random forest regressor

## TOOLS:

- Python version 3.7 (current available)
- Packages : BeautifulSoup, requests, pandas, numpy, matplotlib
- Windows text editor or equivalent software
- Tableau Software
- Microsoft excel or equivalent spreadsheet software
- JupyterLab

# DATA GATHERING AND PREPROCESSING

## Data set after Scrapping

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | brand | model | model_ye | list_price | color | configratic | condition | body_type | wheel_cor | transmissi | fuel_type | mileage | carfax_lin | vin_numbe | image_link | dealer_address | | | | |
| 2 | Acura | TSX | 2006 | 5400 | Blue | w/A-Spec | Used | Sedan | Front-whe | Automatic | Gasoline | 167385 | https://wv | jh4cl968x6 | https://i.el | Old Kennedy Rd, Markham, ON L3R 0L5, Canada | | | | |
| 3 | Nissan | Sentra | 2019 | 16495 | Black | SV \|MOON | Used | Sedan | Front-whe | Automatic | Gasoline | 48567 | https://reports.carpr | | https://i.el | 1599 Star Top Road, Ottawa, ON, K1B 5P5 | | | | |
| 4 | Mercedes- | C-Class | 2018 | 34995 | White | C300\|4MA | Used | Sedan | All-wheel | Automatic | Other | 63851 | https://reports.carpr | | https://i.el | 100 Toro Road, North York, ON, M3J 2A9 | | | | |
| 5 | Honda | Other | 2010 | 4999 | Silver | Sport | Used | Sedan | Front-whe | Manual | Gasoline | 136500 | https://wv | 2HGFA1E6 | https://i.el | North York, ON M6A 2X3 | | | | |
| 6 | Dodge | Grand Car | 2014 | 0 | Silver | 30th ANNI | Used | Minivan, V | Front-whe | Automatic | Other | 157001 | https://wv | 2c4rdgbg6 | https://i.el | 56 Martin Ross Ave., North York, ON, M3J 2L4 | | | | |
| 7 | Toyota | RAV4 | 2021 | 44363 | Red | Limited | New | SUV, Cross | All-wheel | drive (AWD | Gasoline | 0 | | | https://i.el | 2336 Saint Clair Avenue West, Toronto, ON, M6N 1K8 | | | | |
| 8 | Ram | 1500 | 2021 | 79289 | Black | Limited Lo | New | Pickup Tru | 4 x 4 | Automatic | Other | 0 | | | https://i.el | 212 Lakeshore Road West, Mississauga, ON, L5H 1G6 | | | | |
| 9 | BMW | 3-Series | 2013 | 25000 | Black | | Used | Convertible | | | | 140000 | | | https://i.el | Toronto, ON M1B4K4 | | | | |
| 10 | Chevrolet | Cruze | 2014 | 5990 | Blue | LT ~AUTO | Used | Sedan | Front-whe | Automatic | Other | 148000 | https://wv | 1G1PC5SB | https://i.el | 1113 Finch Ave W., Toronto, ON, M3J 2E5 | | | | |

**Figure 3: Data set after cleaning**

## Data set after required modification to be used as input for algorithms

| Car_Name | Year | Selling_Price | Present_Price | Kms_Driven | Fuel_Type | Seller_Type | Transmission | Owner |
|---|---|---|---|---|---|---|---|---|
| ritz | 2014 | 3.35 | 5.59 | 27000 | Petrol | Dealer | Manual | 0 |
| sx4 | 2013 | 4.75 | 9.54 | 43000 | Diesel | Dealer | Manual | 0 |
| ciaz | 2017 | 7.25 | 9.85 | 6900 | Petrol | Dealer | Manual | 0 |
| wagon r | 2011 | 2.85 | 4.15 | 5200 | Petrol | Dealer | Manual | 0 |
| swift | 2014 | 4.6 | 6.87 | 42450 | Diesel | Dealer | Manual | 0 |
| vitara brezza | 2018 | 9.25 | 9.83 | 2071 | Diesel | Dealer | Manual | 0 |
| ciaz | 2015 | 6.75 | 8.12 | 18796 | Petrol | Dealer | Manual | 0 |
| s cross | 2015 | 6.5 | 8.61 | 33429 | Diesel | Dealer | Manual | 0 |
| ciaz | 2016 | 8.75 | 8.89 | 20273 | Diesel | Dealer | Manual | 0 |

**Figure 4: Data set after required modification**

# RESULTS:

We used a car dataset for testing our model. And our findings are:-

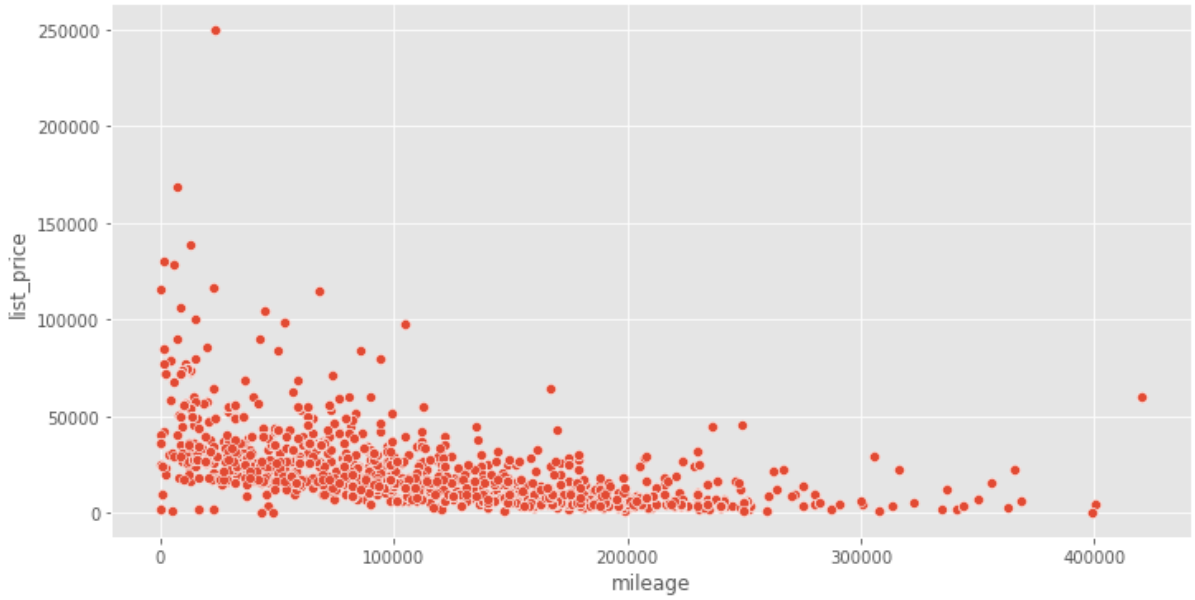



**Figure 5: Regression of price and mileage**

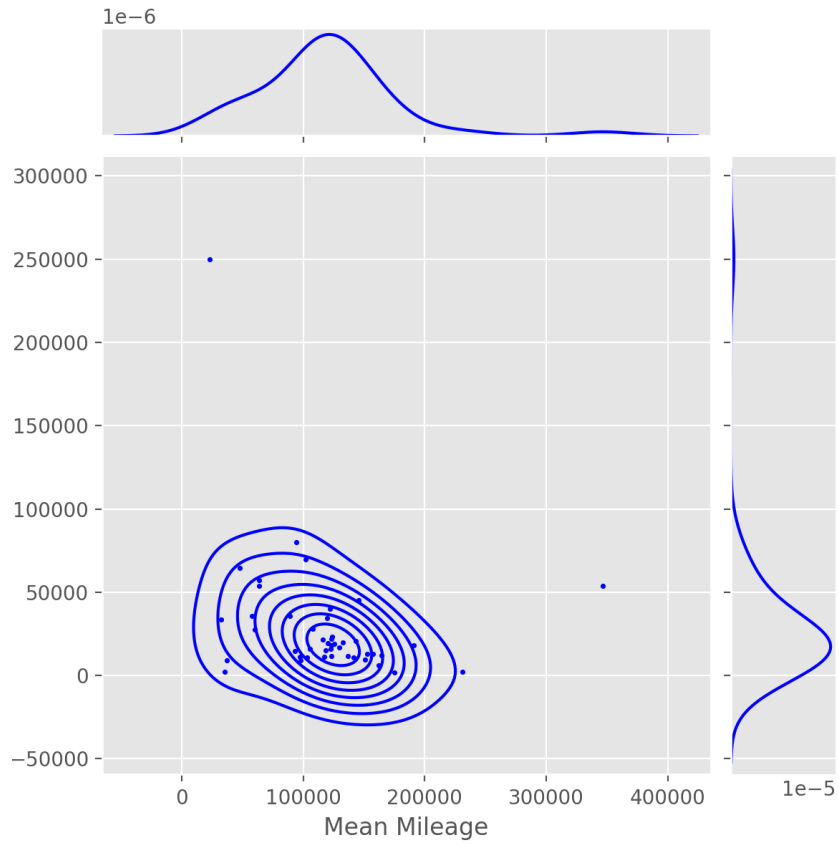



**Figure 6: Regression of mean price – mile**

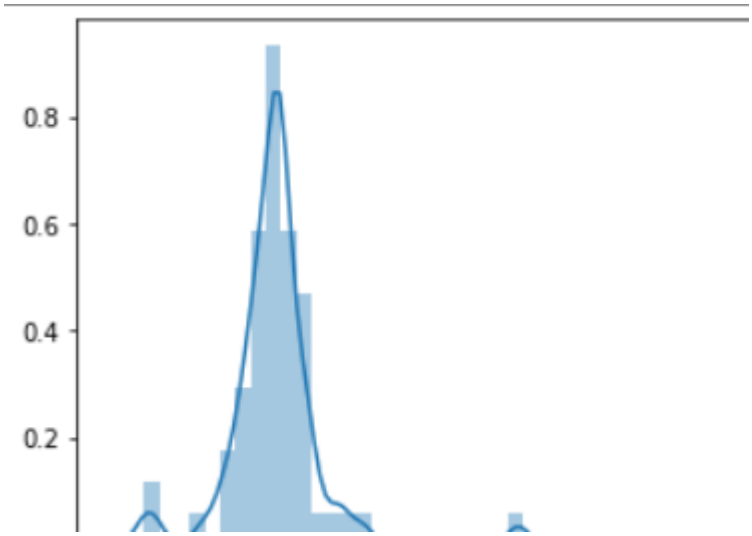



**Figure 7: Correctness in mean price**

# OUTPUT:

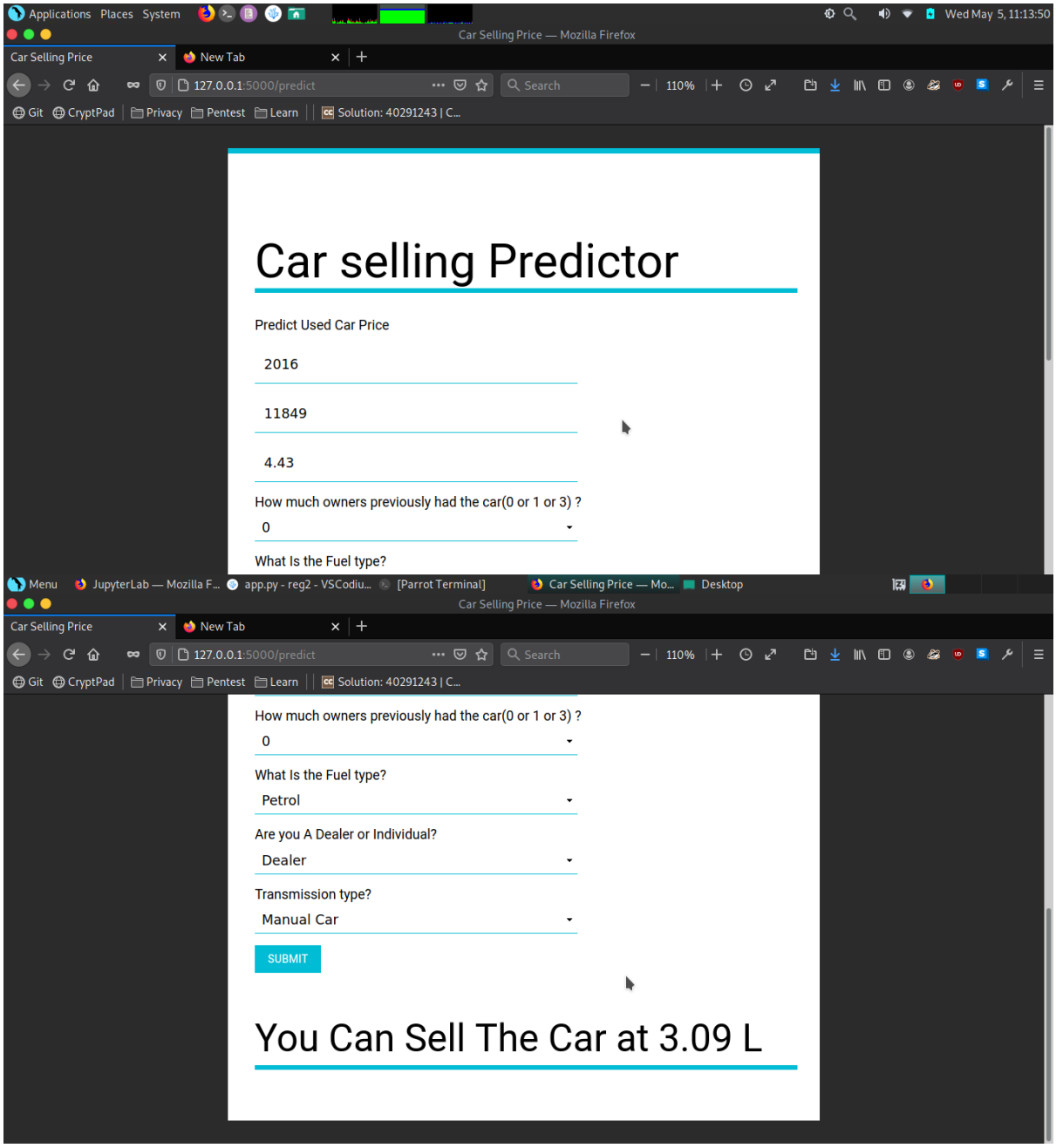


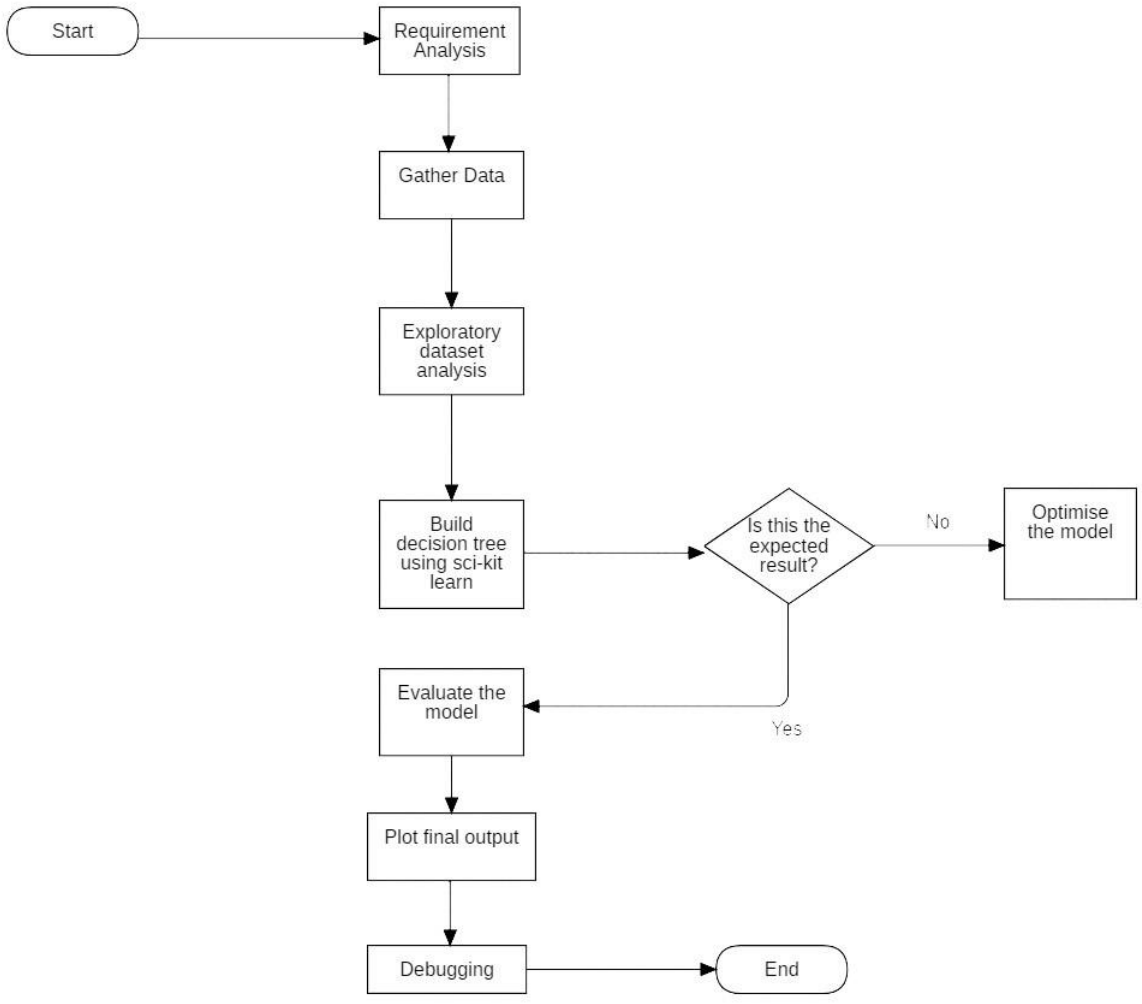

**Figure 8: Prediction of Selling Price after applying regression.**

## SYSTEM REQUIREMENTS:

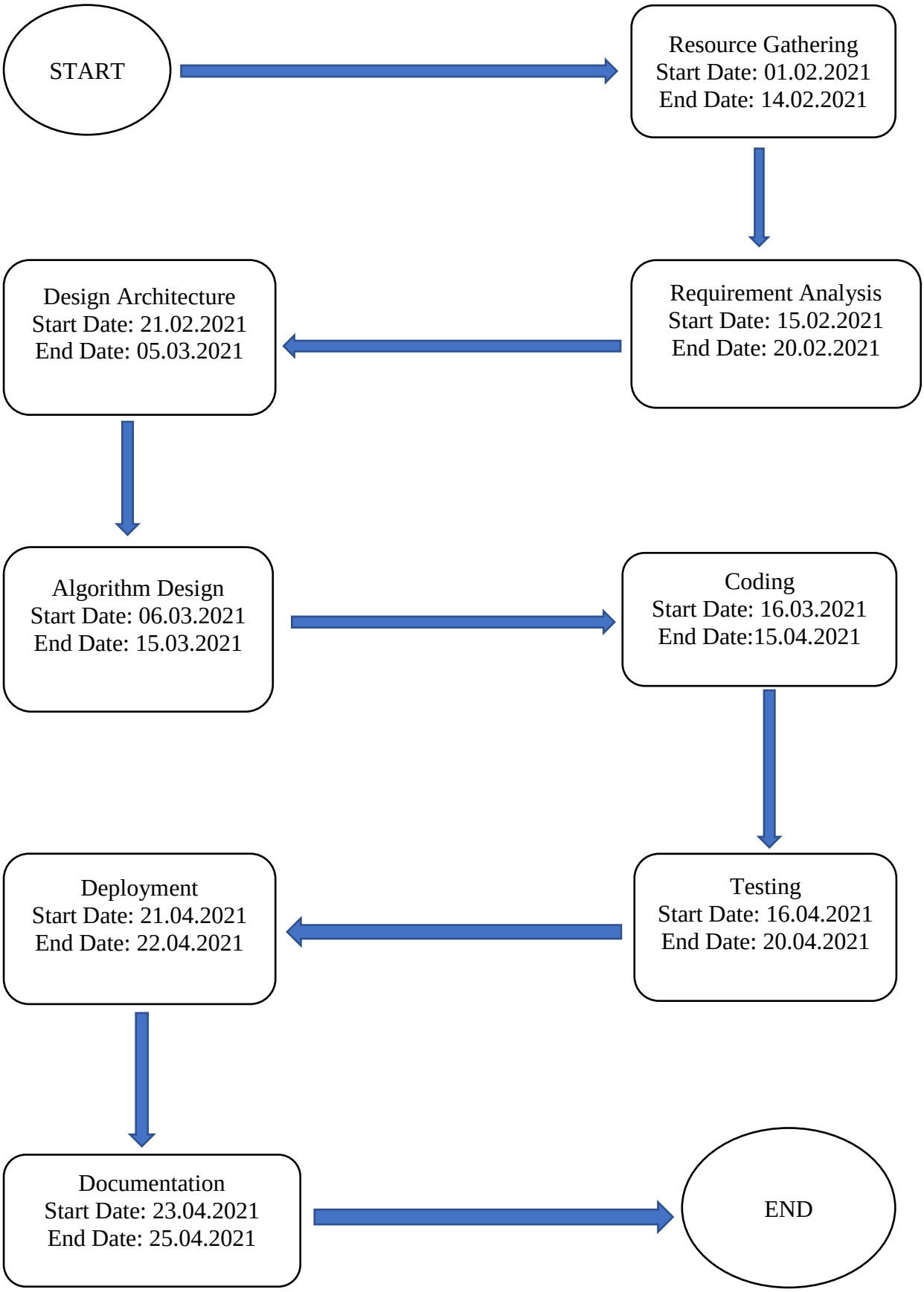
Table 1: Hardware Requirements

| Hardware Components | Configuration |
|---|---|
| Processor | Intel i5 7200 |
| Processor Speed | 2.5 GHz |
| RAM Size | 8 GB |
| OS | Windows 10 |
| GPU | NVIDIA GeForce GTX 940M |
| VRAM | 2GB |

## PROJECT FLOW:



## SCHEDULE (PERT CHART):



## REFERENCES:


*[1]. Agile Methodology & Model: Guide for So ware Development & Tes ng. 1 9.*

*[2]. Data Cleaning: Current Approachesand Issues1Vaishali Chandrakant Wangikar and 2Ratnadeep R. Deshmukh1MCA Department, Maharashtra Academy of Engineering, Alandi,*

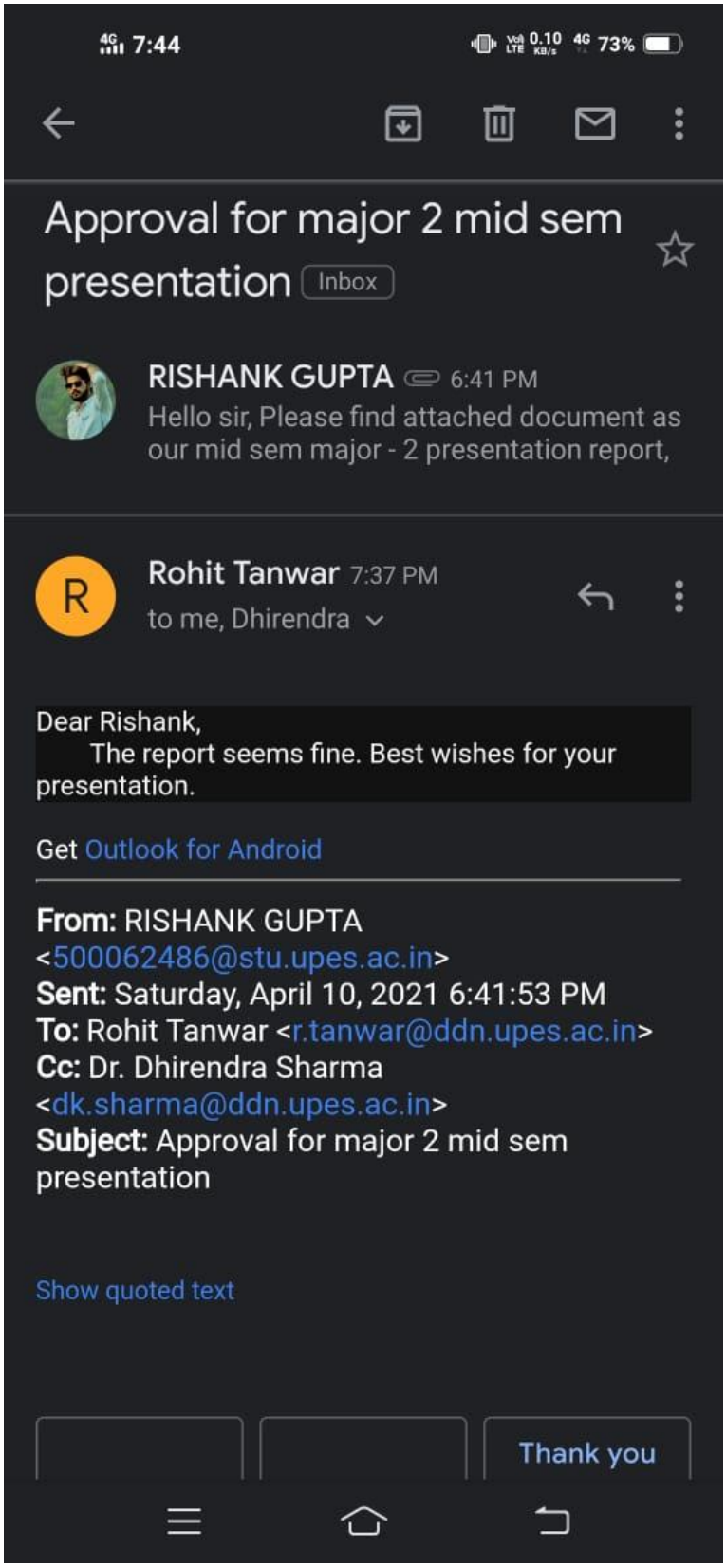
*[3]. Pune (MS), India,2Deptartment of Computer Science & IT, Dr. Babasaheb Ambedkar Marathwada*

*[4]. University, Aurangabad (MS), IndiaE-mail: vaishali.wangikar@gmail.com, ratnadeep_deshmukh@yahoo.co.in*

*[5]. Interna onal Journal for Research in Applied Science & Engineering Technology (IJRASET) ©IJRASET: All Rights are Reserved 335 Web Informa on Retrieval Using Python and Beau fulSoup Pra ksha Ashiwal1, S.R.Tandan2, Priyanka Tripathi3 , Rohit Miri4 Department of Computer Science and Engineering 1,2,4Dr. C V Raman University, Bilaspur, CG, India 3Na onal Ins tute of Technical Teach*

*[6]. Exploi ng web scraping in a collabora ve filteringbased approach to web adver singEloisa Vargiu1, 2, Mirko Urru11. Dipar mento di Matema ca e Informa ca, Università di Cagliari, Italy. 2. Barcelona Digital Technology Centre, SpainCorrespondence: Eloisa Vargiu. Address: Barcelona Digital Technology Center, Italy. Email: evargiu@bdigital.org.*

*[7]. Kabir, Syed Muhammad. (2014). Methods of Data Collec on.*


7:44
73%
Approval for major 2 mid sem presentation
Inbox
RISHANK GUPTA
6:41 PM
Hello sir, Please find attached document as our mid sem major - 2 presentation report,
R
Rohit Tanwar 7:37 PM
to me, Dhirendra
Dear Rishank,
The report seems fine. Best wishes for your presentation.
Get Outlook for Android
From: RISHANK GUPTA <500062486@stu.upes.ac.in>
Sent: Saturday, April 10, 2021 6:41:53 PM
To: Rohit Tanwar <r.tanwar@ddn.upes.ac.in>
Cc: Dr. Dhirendra Sharma <dk.sharma@ddn.upes.ac.in>
Subject: Approval for major 2 mid sem presentation
Show quoted text
Thank you