**A MINI PROJECT REPORT**

**ON**

**“YOUTUBE VIDEO VIRALITY PREDICTION AND ANALYSIS”**

Submitted to

SAVITRIBAI PHULE PUNE UNIVERSITY

in completion of

**SKILL DEVELOPMENT LABORATORY**

**(T.E Computer Engineering)**

**BY**

Name of the student Rhishabh Hattarki Exam No : T150234276

Name of the student Ghoshit Bhelawe Exam No : T150234227



Department of Computer Engineering

Sinhgad College of Engineering, Pune-41

**Accredited by NAAC with grade ‘A’**

**YEAR 2017-18**

**CERTIFICATE**



Sinhgad Technical Education Society,

Department of Computer Engineering

Sinhgad College of Engineering, Pune-41

**Accredited by NAAC with grade ‘A’**

**“YOUTUBE VIDEO VIRALITY PREDICTION AND ANALYSIS”**

Submitted to

SAVITRIBAI PHULE PUNE UNIVERSITY

in completion of

**SKILL DEVELOPMENT LABORATORY**

**(T.E Computer Engineering)**

BY

Name of the student Rhishabh Hattarki Exam No : T150234276

Name of the student Ghoshit Bhelawe Exam No : T150234227

Prof. P. P. Chothe Prof. M. P. Wankhade

Internal Guide Head of Dept.

Department of Computer Engineering Department of Computer Engineering

Dr. S.D. Lokhande

Principal

SCOE, Pune

**CONTENTS**

# TITLE PAGE NO

# 

Certificate I

Acknowledgement II

Abstract III

List of Figures IV

List of Tables V

List of Abbreviation VI

1. **INTRODUCTION** 
   1. Background And Basics
   2. Problem Statement
      1. Scope Statement

**2**. **PROJECT PLANNING & MANAGEMENT**

* 1. Software and Hardware requirement
  2. Process Model

**3. ANALYSIS & DESIGN**

3.1 Use-Case Diagrams

3.2 Class Diagrams

3.3 ER Diagrams(if applicable)

3.4 Sequence Diagrams /DFD & CFDs

3.5 State Transition Diagrams

3.6 Deployment Diagrams

3.7 Component Diagram

**4. IMPLEMENTION & CODING**

* 1. Methodology
  2. Dataset

4.2 Algorithms/Flow Charts

4.3 GUI Design /screen shots

**5. RESULTS & DISCUSSION**

4.1 Visualization of results (Graphs, Charts, etc.)

**CONCLUSION**

**REFERENCES**

**Acknowledgement**

We would like to express our special thanks of gratitude to our teacher Prof. P. P. Chothe as well as our college Sinhgad College of Engineering who gave us the golden opportunity to do this wonderful project on the topic Youtube Video Virality Prediction and Analysis, which also helped us in doing a lot of Research and we came to know about so many new things we really thankful to them. We would also like to thank Prof. Shakeeb Sheikh for explaining concepts necessary for the project.

**Abstract**

YouTube**,** LLC is an American video-sharing that allows users to upload, view, rate, share, add to favorites, report, comment on videos, and subscribe to other users.

YouTubers (users uploading content on YouTube) earn advertising revenue from Google AdSense, a program which targets ads according to site content and audience. YouTube connects the channels with the advertisers based on factors like video-viewer engagement. This includes views, likes, dislikes, comments.

The report documents the prediction of one of the above given factors – video likes, that allows a YouTuber maximize their revenue. It uses machine learning technique – Random Forest Regression for the prediction. Data analysis has been done that provides useful insights.

**List of Figures**

Fig 4.4.1 PyQt5 designer – used to create the ui

Fig 4.4.2 Homepage of GUI

Fig 4.4.3 ML – Prediction Page

Fig 4.4.4 Analysis page

Fig 5.1.1 Comments vs Likes – scatter graph, predicted values on test data – plot graph

Fig 5.1.2 Word Count

Fig 5.1.3 Countries in absolute numbers

**List of tables**

Table 4.2.1 Dataset

* + - 1. **INTRODUCTION**
  1. **BACKGROUND AND BASICS**

**1.1.1 YouTube - Introduction**

As of February 2017, there were more than 400 hours of content uploaded to YouTube each minute, and one billion hours of content being watched on YouTube every day. As of August 2018, the website is ranked as the second-most popular site in the world, according to Alexa Internet.

In May 2007, YouTube launched its Partner Program (YPP), a system based on AdSense which allows the uploader of the video to share the revenue produced by advertising on the site. YouTube typically takes 45 percent of the advertising revenue from videos in the Partner Program, with 55 percent going to the uploader. On January 16, 2018, the eligibility requirement for monetization was changed to 4,000 hours of watchtime within the past 12 months and 1,000 subscribers.

**1.1.2 Machine Learning**

Machinelearning is a field of artificial intelligence that uses statistical techniques to give computer systems the ability to "learn" (e.g., progressively improve performance on a specific task) from data, without being explicitly programmed. Within the field of data analytics, machine learning is a method used to devise complex models and algorithms that lend themselves to prediction; in commercial use, this is known as predictive analytics. These analytical models allow researchers, data scientists, engineers, and analysts to "produce reliable, repeatable decisions and results" and uncover "hidden insights" through learning from historical relationships and trends in the data.

Types of Machine Learning Techniques:

* Supervised Learning
  + Regression
    - Linear Regression
    - Random Forest Regression
    - SVR
    - Decision Trees
  + Classification
    - Support Vector Machines
    - Naïve Bayes
    - Nearest Neighbour
* Unsupervised Learning
  + Clustering
    - K-means
    - Heirarchical
    - Gaussian Mixture
  + Dimension Reduction
    - PCA
* Reinforcement Learning

**1.1.3 Data Analysis**

Dataanalysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, while being used in different business, science, and social science domains. Analysis refers to breaking a whole into its separate components for individual examination. Data analysis is a process for obtaining raw data and converting it into information useful for decision-making by users. Data is collected and analyzed to answer questions, test hypotheses or disprove theories.

* 1. **PROBLEM STATEMENT**
     1. **Scope statement**

The monetization of a YouTube video depends on the video engagement. More the video engagement, more is the revenue earned by the YouTuber. The video engagement can be described as a cumulation of the number of views earned by the video, number of likes, number of comments, number of dislikes. Increasing these figures of a video would benefit the uploader. As a result predicting these values can help an uploader understand the range and scope of their video.

With the increasing number of videos, the amount of data is increasing at an unprecedented rate. Without understanding the data properly, it is of no use. Hence, data analysis is necessary. Data analysis on the YouTube dataset provides an insight on the different aspects of videos in different countries that get to the trending page.

* + - 1. **PROJECT PLANNING AND MANAGEMENT**

**2.1 SOFTWARE REQUIREMENTS**

**2.1.1 Python 3.7 64 bit**

**2.1.2 Spyder 3**

**2.1.3 PyQt Designer 5**

**3. ANALYSIS AND DESIGN**

**3.1. Use Case Diagram**

A use case diagram is a **dynamic** or behavior diagram in UML. Use case diagrams model the functionality of a system using actors and use cases. Use cases are a set of actions, services, and functions that the system needs to perform.

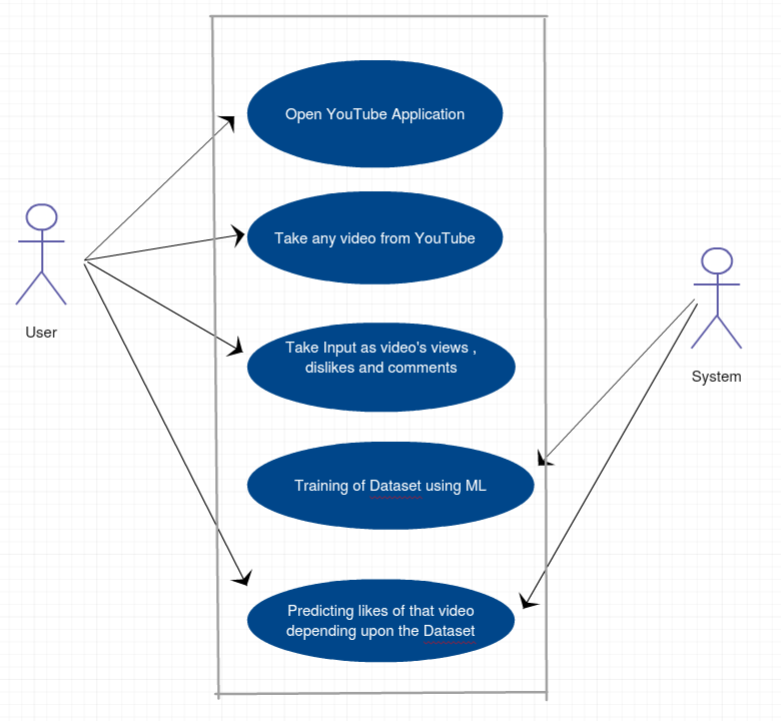


Fig 3.1 Use Case Diagram

**3.2. Sequence Diagram**

A sequence diagram simply depicts interaction between objects in a sequential order i.e. the order in which these interactions take place .Sequence diagrams describe how and in what order the objects in a system function.

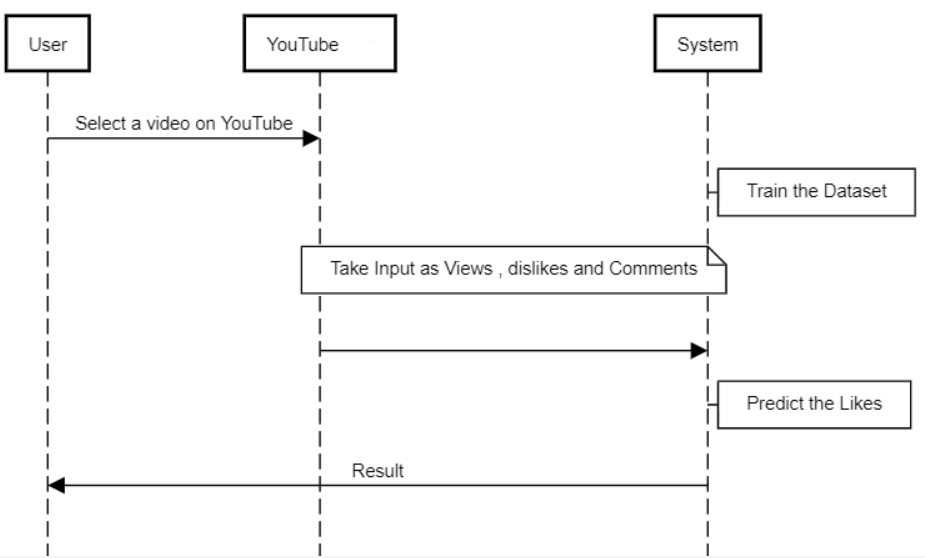
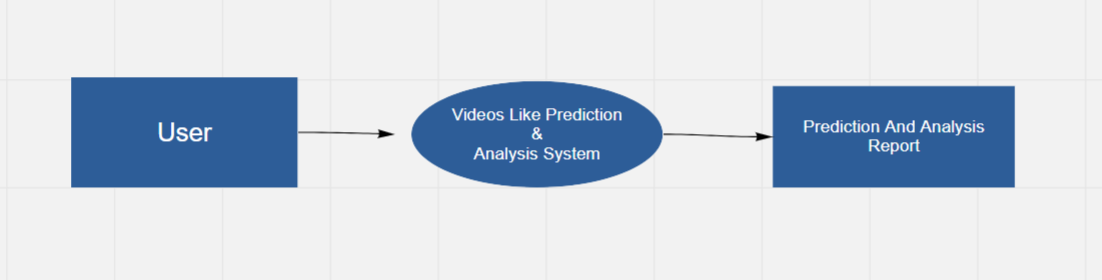


Fig 3.2 Sequence Diagram

**3.3. Data Flow Diagram**

A **data flow diagram** (**DFD**) is a graphical representation of the "**flow**" of **data** through an information system, modeling its process aspects. A**DFD** is often used as a preliminary step to create an overview of the system without going into great detail, which can later be elaborated



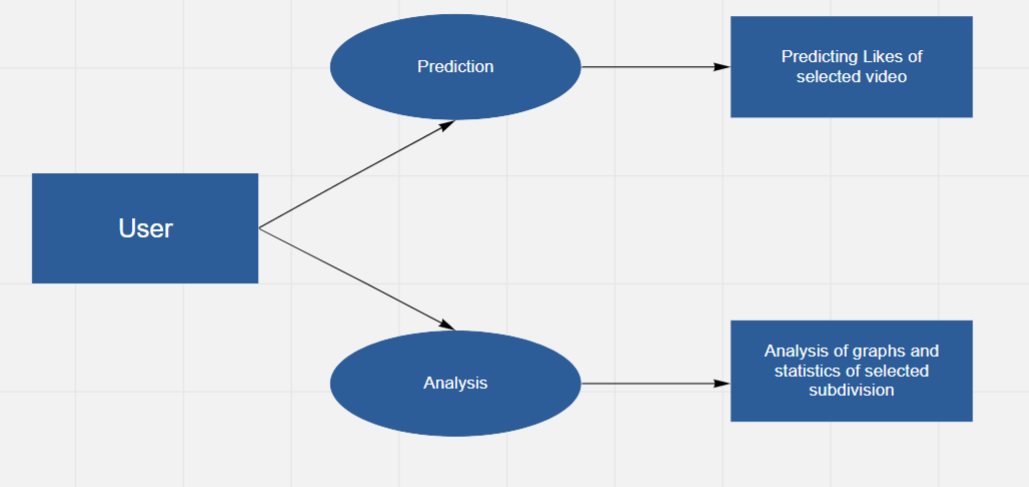


Fig 3.3 DFD

**4. IMPLEMENTATION & CODING**

**4.1 METHODOLOGY**

**4.1.1 Gathering data**

Kaggle is the world's largest community of data scientists and machine learners. It offers various different datasets to work with. The dataset used in this project has been taken from kaggle. It consists of data of 5 different countries. The datapoints describe the videos that have reached the trending page.

**4.1.2 Data Preparation**

Since the data was collected from 5 different countries consisting of 40k datapoints each, it was combined into one single dataset of 200k datapoints. The machine learning model was trained on this combined dataset.

Some columns were not the required format that needed to be converted. For example the date and time columns, in the data analysis part of calculating days required to reach trending page, the datetime column had to converted into a format easily handled by pandas. The date time column had to be split into separate columns.

The dataset is split into dependent and independent variables' columns. They are basically inputs and outputs given to the model. This data is further split into training and testing data in the ratio 80-20.

**4.1.3 Choosing the model**

Among the multiple choices available in the sklearn library for choosing the model, we tested Random Forest Regression, Linear Regression, Ridge, KNeighbors Regression, Decision Tree Regression, MLP Regression, Gradient Boost Regression.

The models are added in a list. Using a technique called k fold cross validation, the models are tested. The training data is further split into 7 parts and in 7 iteration, using every part as test data per iteration, the accuracy scores are aquired. These scores are further averaged to get a better result.

The accuracies of the above models were as follows:

Random Forest - 98.178

Linear Regression - 85.568

Ridge - 85.568

KNN Regression - 80.789

Decision tree - 97.300

MLP Regression - 85.301  
Gradient boost - 94.284

Since the highest accuracy was achieved by Random Forest, this was chosen as the preferred model.

**4.1.4 Training**

The data was split into train test split like in the previous step into training and test data in ratio 80-20. The model is fit to the training data.

**4.1.5 Evaluation**

To check whether the used model is accurate, predict function is used on the test data. The accuracy score is measured with the function r2 score.

The accuracy on the test data came out to be - 98.172

**4.1.6 Prediction**

The values required as input from users are number of views, dislikes, comments which are accepted in the GUI. The number of likes are predicted from the pretrained model that has been pickled for efficiency. Pickling is storing the trained model in a file in the binary format, so that we don't need to retrain the model for every prediction.

**4.2 DATASET**

The entire dataset “youtube-new” consists of 5 separate files per country – US (United States), CA (Canada), GB (Great Britain), DE (Deutschland/Germany ), FR (France).

Every country’s dataset contains the following colums –

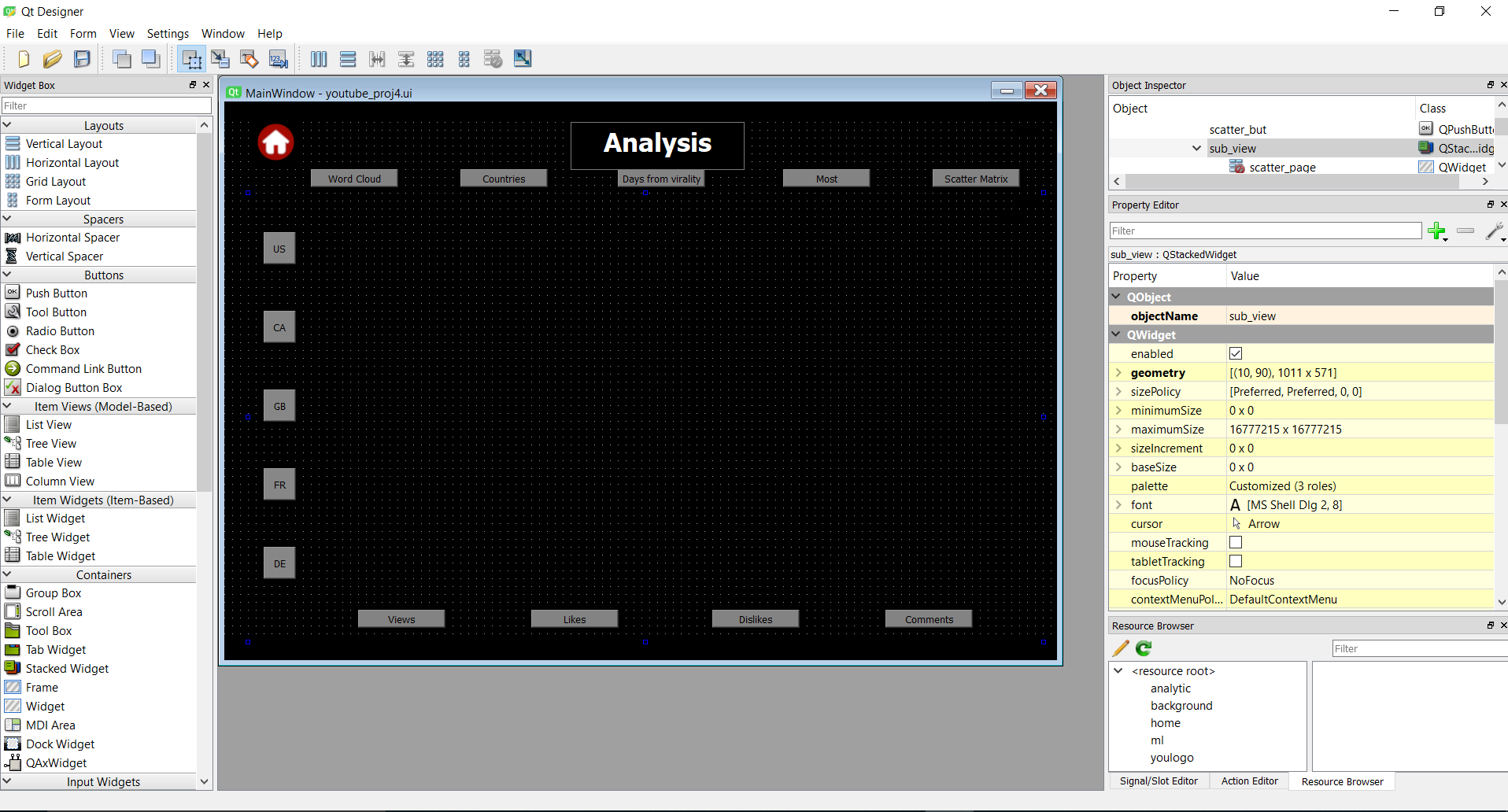
1. Video id
2. Trending date
3. Title
4. Channel title
5. Category id
6. Publish time
7. Tags
8. Tags
9. Views
10. Likes
11. Dislikes
12. Comment count
13. Thumbnail
14. Comments disabled
15. Ratings disabled
16. Video error or removed
17. Description



**Table 4.2.1 Dataset**

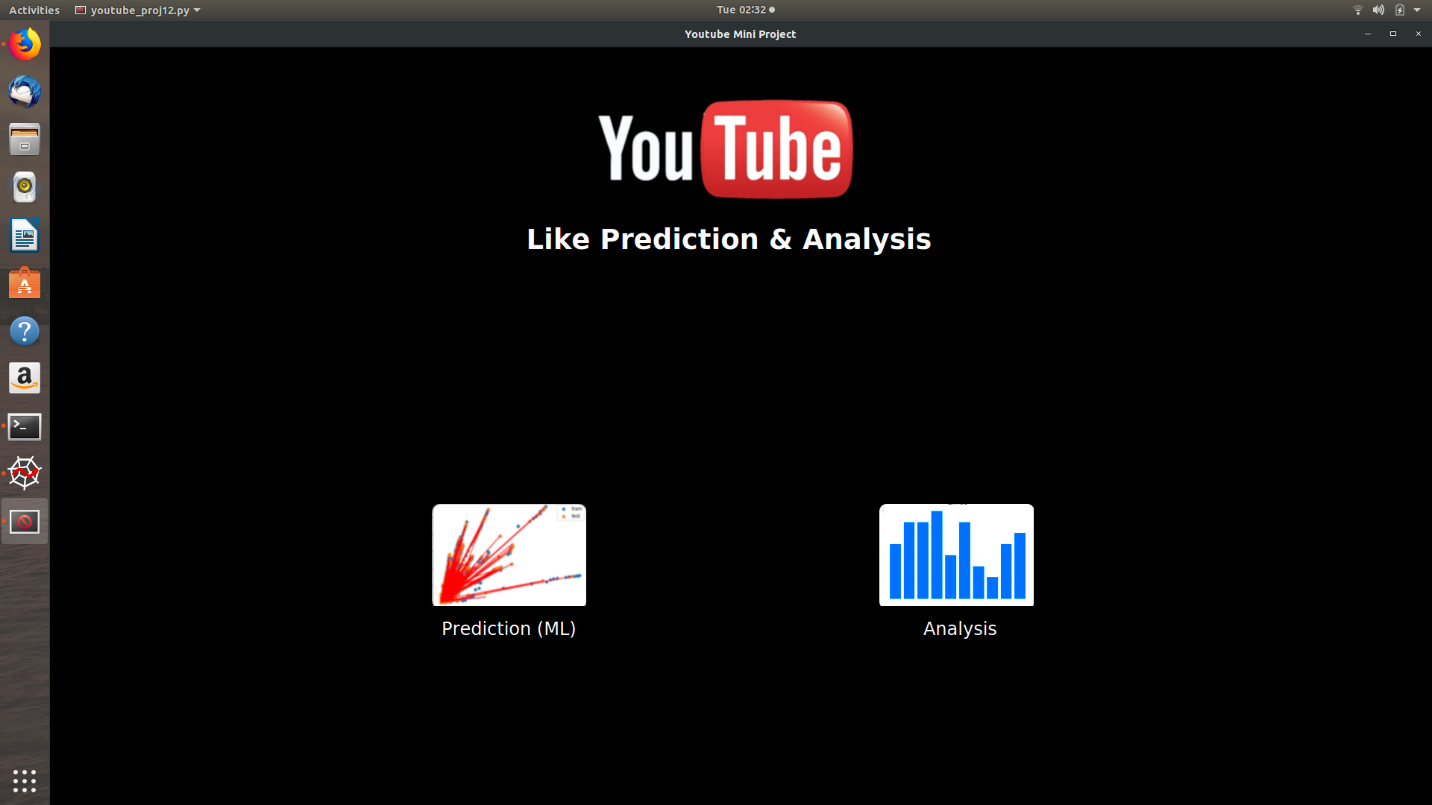
**4.4 GUI DESIGN AND SCREENSHOTS**

The GUI has been designed with the python library PyQt5 and the software PyQt5 designer. The designer is used to create the ui. This ui is then converted into python file using the command – pyuic5 -x miniproj.ui -o miniproj.py. The images required for the project are stored in a qrc file. It needs to be converted into py file and is to be imported in the main program. It can be done using the command – pyrcc miniproj.qrc -o miniproj\_rc.py.



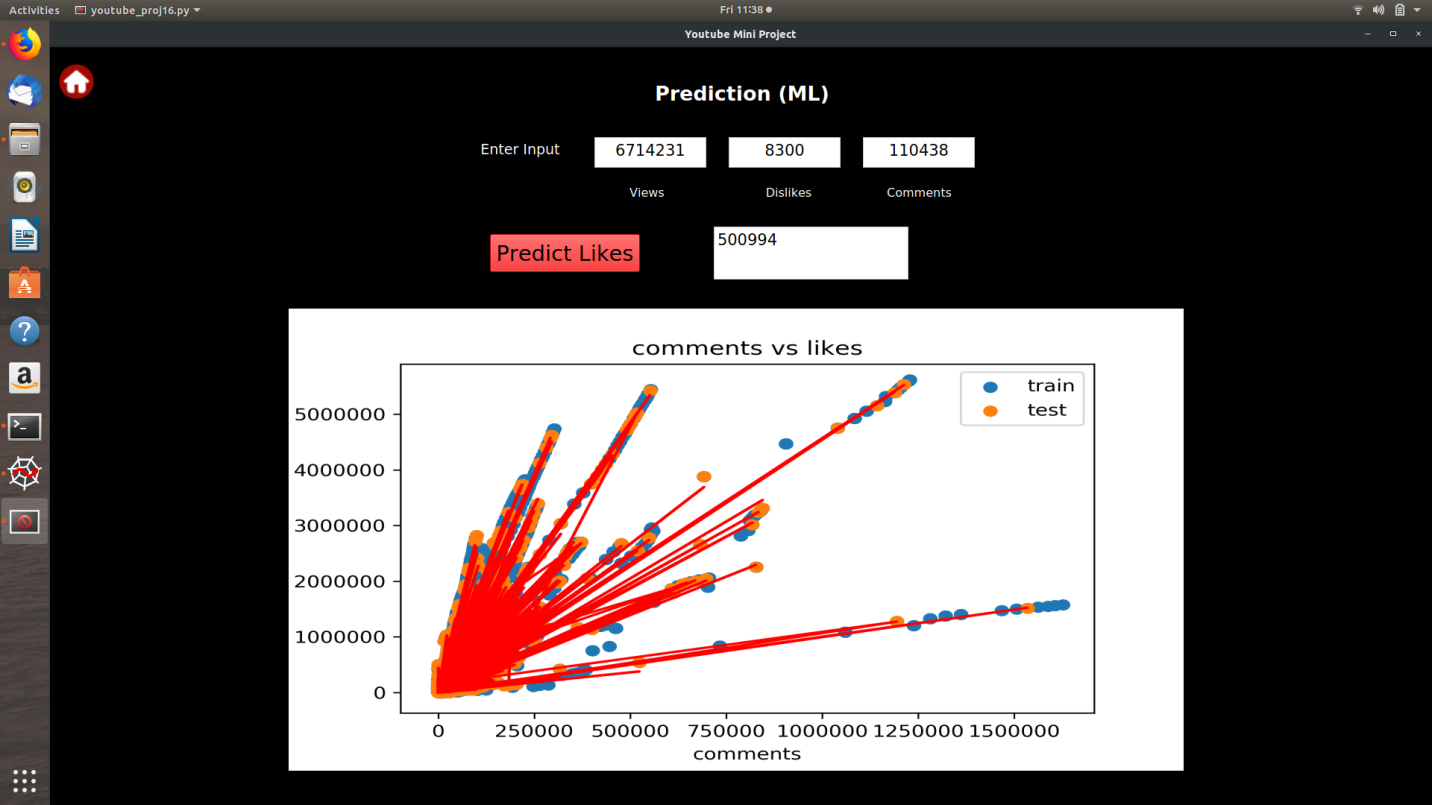
**Fig 4.4.1 PyQt5 designer – used to create the ui**

Above image shows the screenshot of PyQt5 designer. All the pages were created in the designer and the overall layout was set in it. The buttons, text boxes, images were added in it. This is a convenient way of creating a ui without actually coding much, most of the code is added later after converting the files to .py, where the buttons are linked to its respective functions.



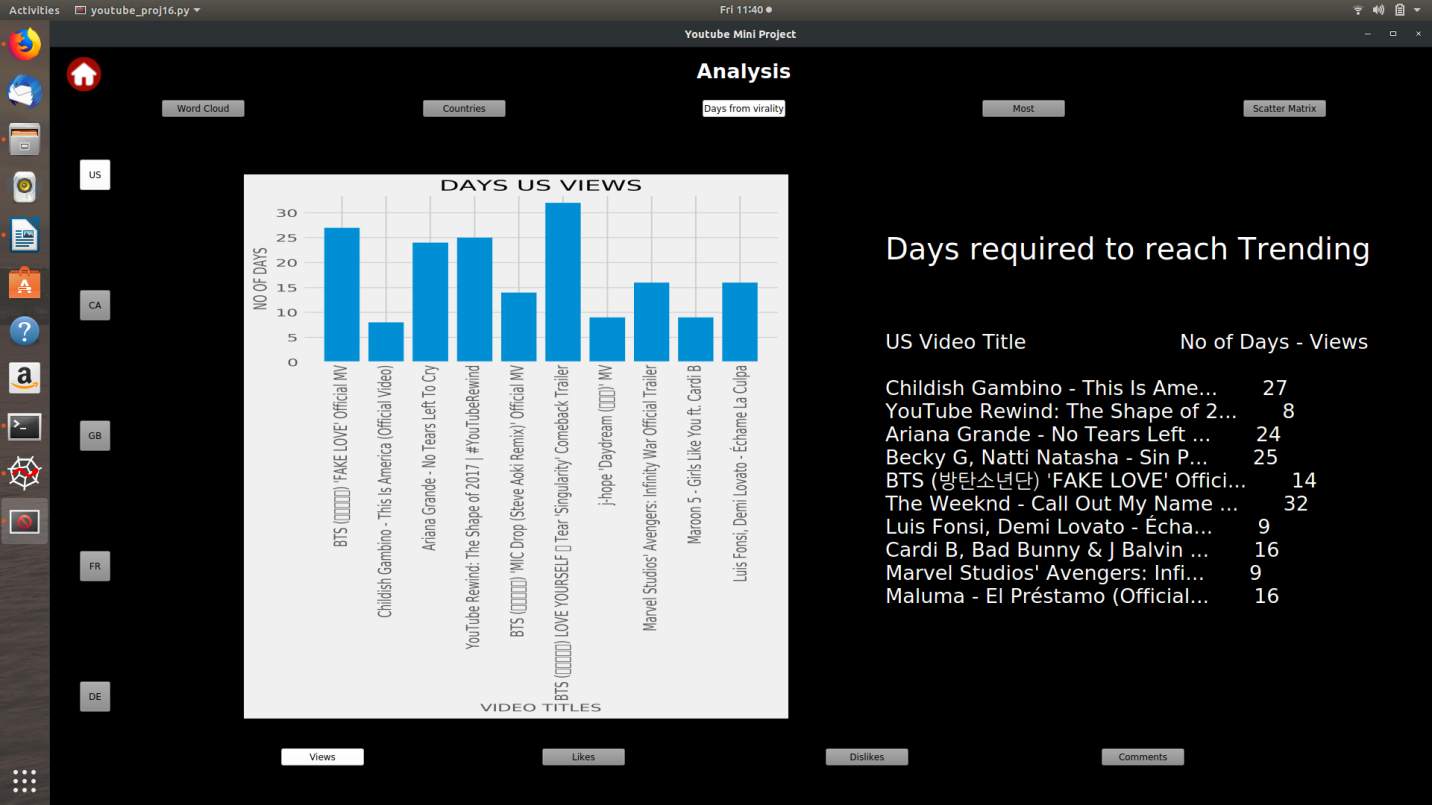
**Fig 4.4.2 Homepage of GUI**

The above screenshot shows the homepage of the project GUI. The prediction button takes the user to the prediction page. The analysis button takes the user to the analysis.



**Fig 4.4.3 ML – Prediction Page**

The above screenshot shows the ML prediction page. The user should input the number of views, dislikes and comments. After clicking on the “Predict likes” button, the predicted number of likes is printed in the text browser.



**Fig 4.4.4 Analysis page**

The above screenshot shows the analysis page. The top 5 buttons change between the type of graphs –

1. Word count
2. Countries in absolute numbers
3. Days from virality
4. Most popular videos
5. Scatter Matrix

The buttons at the side choose between the countries –

1. United States
2. Canada
3. Great Britain
4. France
5. Germany (Deutschland)

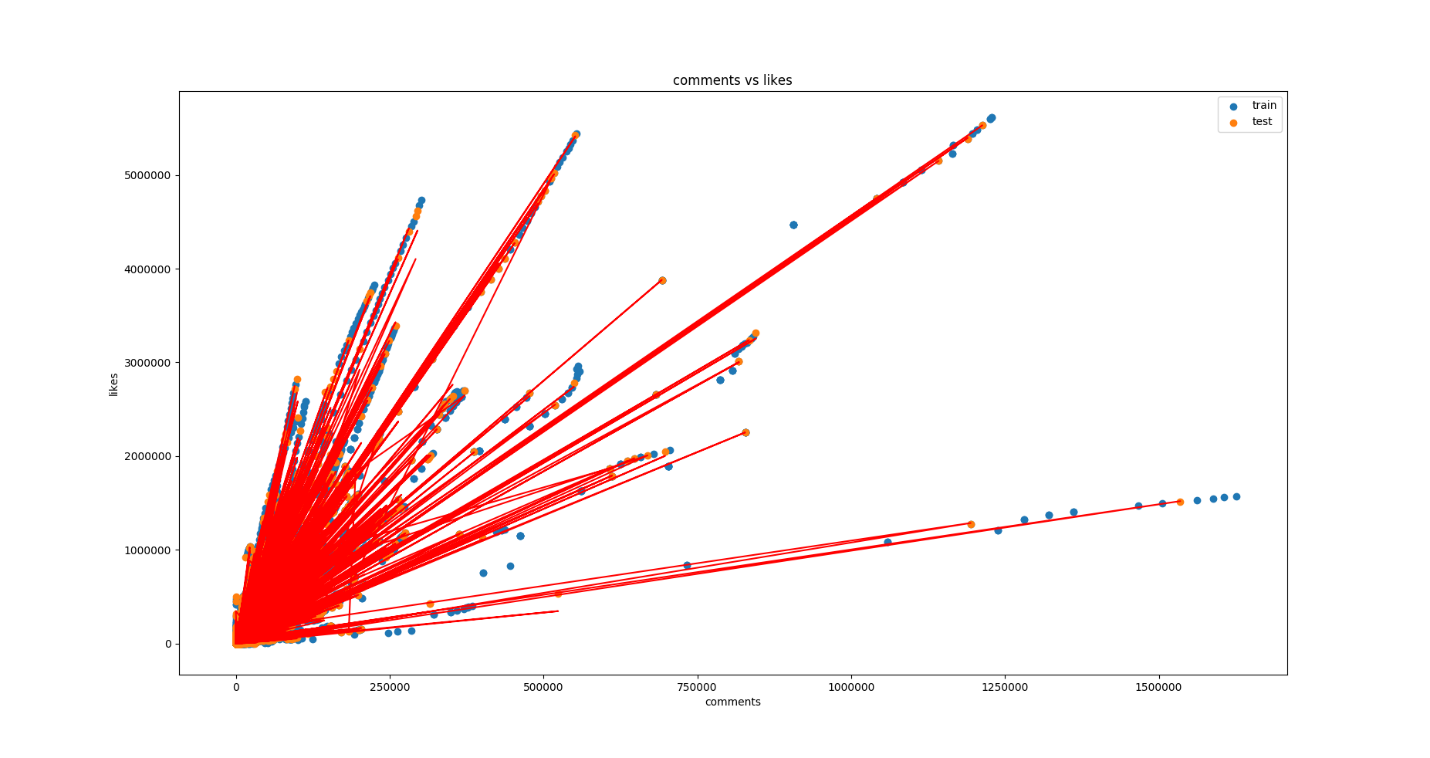
The buttons on the bottom choose between the category –

1. Views
2. Likes
3. Dislikes
4. Comments

**5. RESULTS & DISCUSSIONS**

**5.1 VISUALIZATION OF RESULTS**

Following are few of the graphs that have been plot for visualization and better understanding of the data.

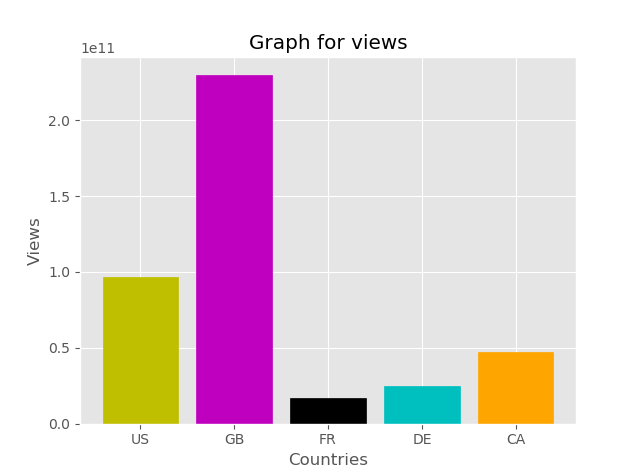


**Fig 5.1.1 Comments vs Likes – scatter graph, predicted values on test data – plot graph**

The above screenshot represents the graph of number of comments on x axis and number of likes on the y axis. The scatter graph in blue shows the training data. The scatter graph in orange shows the test data. The lines plotted in red show the values predicted by random forest regression.



**Fig 5.1.2 Word Cloud**

The image above shows the word cloud made by the entire dataset. A word cloud is an image composed of words used in a particular text or subject, in which the size of each word indicates its frequency or importance.

**Fig 5.1.3 Countries in absolute numbers**

The graph above shows the bars of maximum number of cumulative views per country. The GUI consists of category options where graphs as per likes, dislikes and comments can be viewed.

**CONCLUSION**

Thus, dataset was used to predict number of likes based on number of views, dislikes, comments. The random forest regressor achieved the highest accuracy of 98% on the test data set separated from the same data among other regression algorithms. The analysis of the data provides a better understanding of the data and can be used to achieve better results on video monetization.

**REFERENCES**

* + **World Wide Web**

1. Kaggle, Dataset, <https://www.kaggle.com/datasnaek/youtube-new>
2. Machine Learning mastery, ML tutorial, <https://machinelearningmastery.com/>
3. Sentdex, ML tutorial, <https://www.youtube.com/watch?v=OGxgnH8y2NM>
4. Sentdex, PyQt tutorial, <https://www.youtube.com/watch?v=JBME1ZyHiP8>
5. Allen Wang, Project Reference, <https://towardsdatascience.com/youtube-views-predictor-9ec573090acb>
6. Ayush, Project Reference, <https://github.com/ayush1997/YouTube-Like-predictor/blob/master/report.pdf>
7. Udemy, ML tutorial, <https://www.udemy.com/>