

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | |
|  | DATA ANALYSIS ASSIGNMENT REPORT | | | | |  | |
|  |  | | | | | |  |
|  | |  | | |  | | |
|  | | USAGE OF SOCIAL MEDIA USERS | | |  | | |
|  | | https://tse2.mm.bing.net/th?id=OIP.ilXdk2SXQHe_INbp6BDXogHaFQ&pid=Api&P=0&w=238&h=169 Course by Assoc. Prof. Nguyen Thi Thuy Loan **ĐÀO ANH MINH NGUYỄN MINH TOÀN** | | |  | | |
|  | | |  |  | | | |

Table of Contents

[I.INTRODUCTION 1](#_Toc91716934)

[a) Abstract 1](#_Toc91716935)

[b)The data extraction procedure 2](#_Toc91716936)

[c) The methodologies and tools utilized 4](#_Toc91716937)

[II. PROJECT'S TASK TIMELINE 5](#_Toc91716938)

[III.EXPLORATORY DATA ANALYSIS 6](#_Toc91716939)

[a)Data overview 6](#_Toc91716940)

[b)Handle missing value 7](#_Toc91716941)

[c)Build a predictive model 9](#_Toc91716942)

[IV. CONCLUSION 12](#_Toc91716945)

[Difficulties and mistakes that may impair detection 15](#_Toc91716946)

[Predict 17](#_Toc91716947)

[V.REFERENCES 19](#_Toc91716948)

# I.INTRODUCTION

## a) Abstract

To begin, we must acknowledge the value of advertising.

* **Advertising** is a marketing communication that employs an openly sponsored, non-personal message to promote or sell a product, service, or idea.  Sponsors of advertising are typically businesses wishing to promote their products or services.



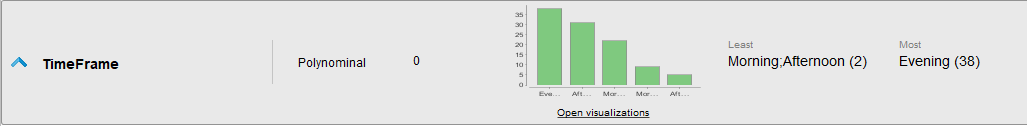
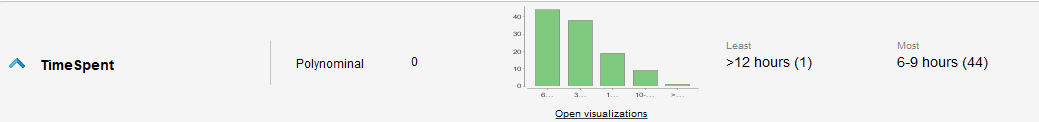
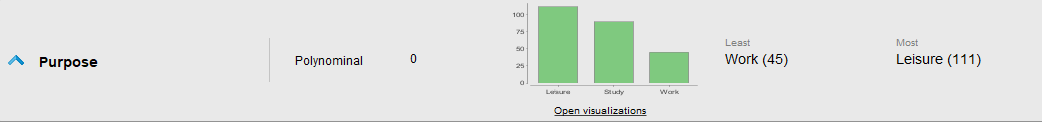
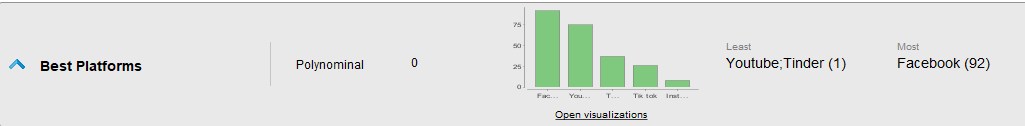
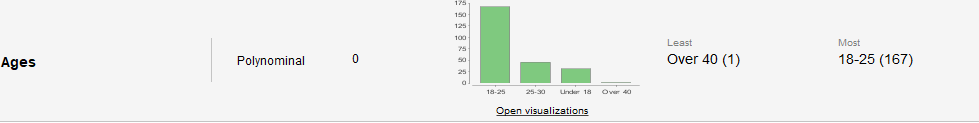
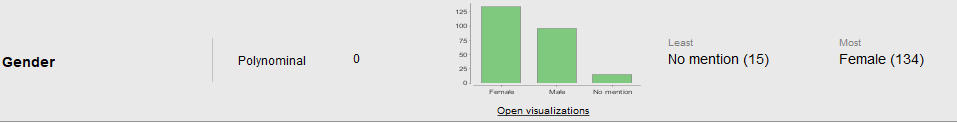
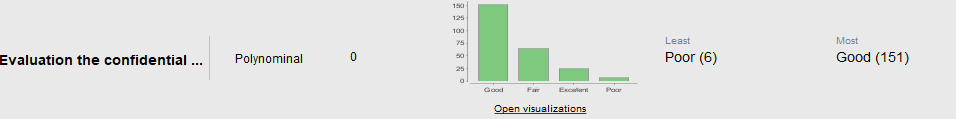
<https://www.webcodebuddy.com/256-step-away-from-the-print-ad-creative-advertising-ideas-thatll-get-noticed/>

* Nowadays, social networks are expanding at a breakneck pace. Each year, the number of visitors to social networking sites grows significantly. As a result, social media is an excellent platform for businesses and suppliers to market their wares.
* Thus, practically every business or supplier has a department called "*marketing*" that is responsible for promoting their products and services. And today, rather than concentrating on things outside of life, such as organizing seminars, fairs, or other events to promote their products, they prefer to bring all of those things to social media. It enables them to save money, time, and effort, but their audience is several times larger than offline.
* However, to efficiently market online, rather than developing a distinct website, we may use existing platforms such as *Facebook, TikTok, YouTube*, and *Instagram*. Additionally, these platforms assist businesses and sellers in advertising their items for a charge. Since the user's purpose of use is so diverse, gathering information about users' social network use patterns would enable the organization to save far more money comparison to entrusting it entirely to social media platforms.
* To gather information on social network users, we can't collect data on everyone who uses social media; instead, we utilize Facebook groups to collect data on a sample of users.

## b)**The data extraction procedure**

* **Collect data**: gathering from many platforms (Facebook, Instagram,…)
* **Select data:** Retrieving data excel and sorting out attributes.
* **Analyze data:** separating groups and predicting statistical evaluation.
* **Visualize data**: Showing predicted data visualization and list of performances.
* <https://forms.gle/dHVd67bS4Le8xCX26>

***Data included:***

* **Timeframe**: Easily determine the time period during which people are most active on social media to assist suppliers and businesses in determining the most cost-effective times to run advertising. (Morning, afternoon, evening ,...) 
* **Timespent**: Also with using timeframes to target a certain set of people throughout time, Timespent assists us in determining an appropriate time period for publishing in order to reach the most audience. (1h-3h, 3h-6h, 6h-9h, 9h-12h ,…) 
* **Purpose**: Understanding between different sorts of things and their uses while utilizing social media. (Study, Work, Entertainment ,... 
* **Platforms**: Inform us of the platform that is most often utilized for each use reason. (Facebook, Telegram, Instagram ,...) 
* **Ages**: Allow us to determine the age group that utilizes social media the most, allowing us to analyze the approach to each age category. 
* **Gender**: Shows that the user's behavior is gender-specific. 
* **Activities**: There are several methods for inserting advertisements. Activities enables us to ascertain the sorts of activities that consumers like and then focus our efforts on those areas.
* **Media**: Enables us to determine which forms of media are most often accessed by consumers, allowing us to target them with advertisements.
* **Evaluation**: User evaluations of the security of social networking sites will provide us with a wealth of information about which platforms are the most secure and popular. 

## c) The methodologies and tools utilized

* ***RapidMiner*** was used to perform all pre-processing, analysis, and visualization of the datasets through charts.
* Using ***Google Forms*** to collect raw data.
* ***Microsoft Excel*** for handling values that are not present (some of the steps).

# II. PROJECT'S TASK TIMELINE

|  |  |  |
| --- | --- | --- |
| **Part** | **Activity** | **MEMBERS** |
| **INTRODUCTION** | Research Advertising definition | Đào Minh |
| The data extraction procedure | Minh Toàn |
| Making Google Forms survey | Both |
| Topic confirmation | Both |
| Determine the goal and purpose of the analysis | Minh Toàn |
| **OVERVIEW** | General analysis and viewing the data by using Rapidminer | Minh Toàn |
| Report data overview and handle missing values | Đào Minh |
| Group discussion | Both |
| **VISUALIZING** | Generated appropriate charts on criteria | Đào Minh |
| Explain the charts and visualization attributes | Minh Toàn |
| Using predictive model in Rapidminer | Đào Minh |
| Group discussion | Both |
| **PRESENTATION** | Make the report | Both |
| Prepare for the slides | Both |
| Prepare the script for better presentation | Both |
| Practice presentation | Both |

# III.EXPLORATORY DATA ANALYSIS

## a)Data overview

* Begin by collecting raw data using GG forms and then exporting to file.csv for input into *Rapidminer*. From there, utilize the **Data, Statistics, and Visualization** tab to get the a forementioned findings.

***RAW DATA***

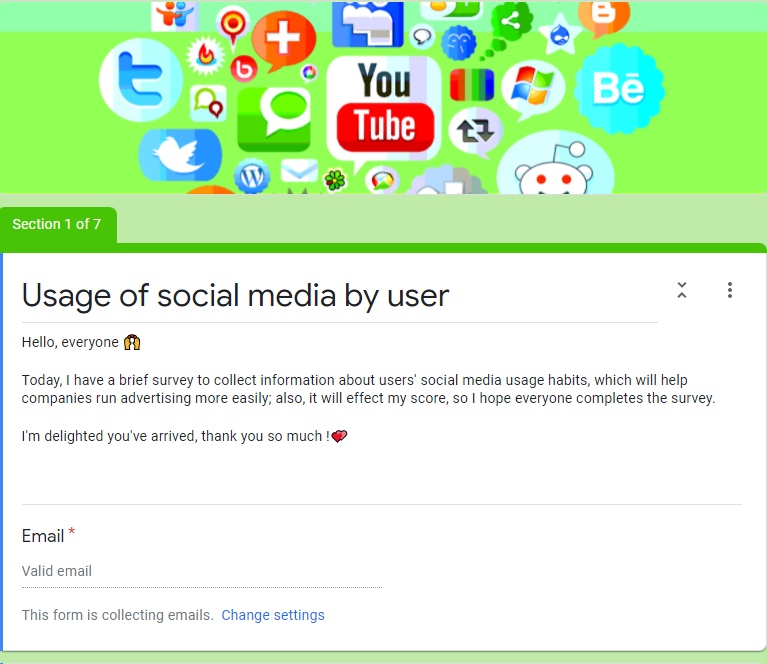
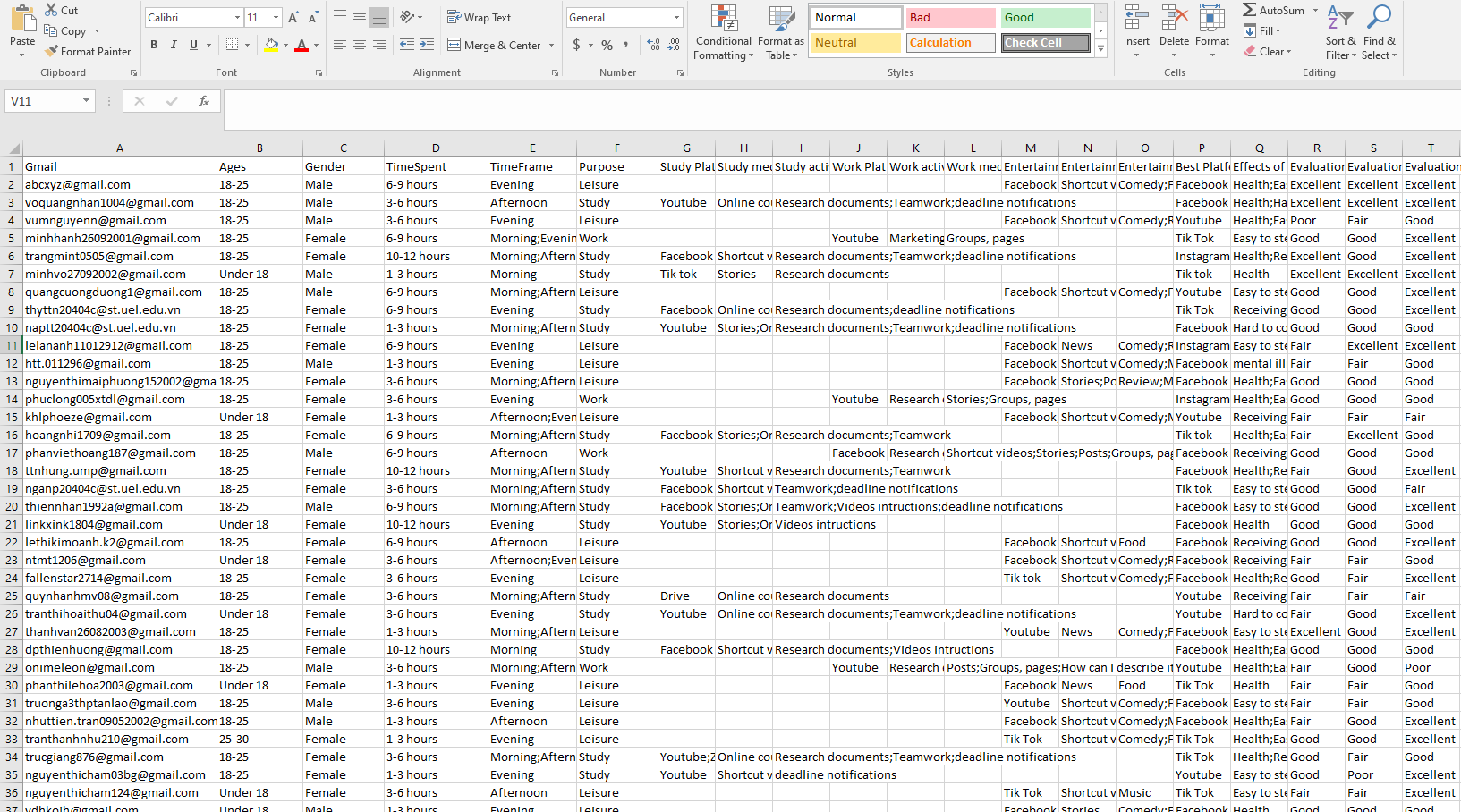
 

Figure 1: GG FORMS collecting data Figure 2: FIle.csv

***RAPIDMINER***

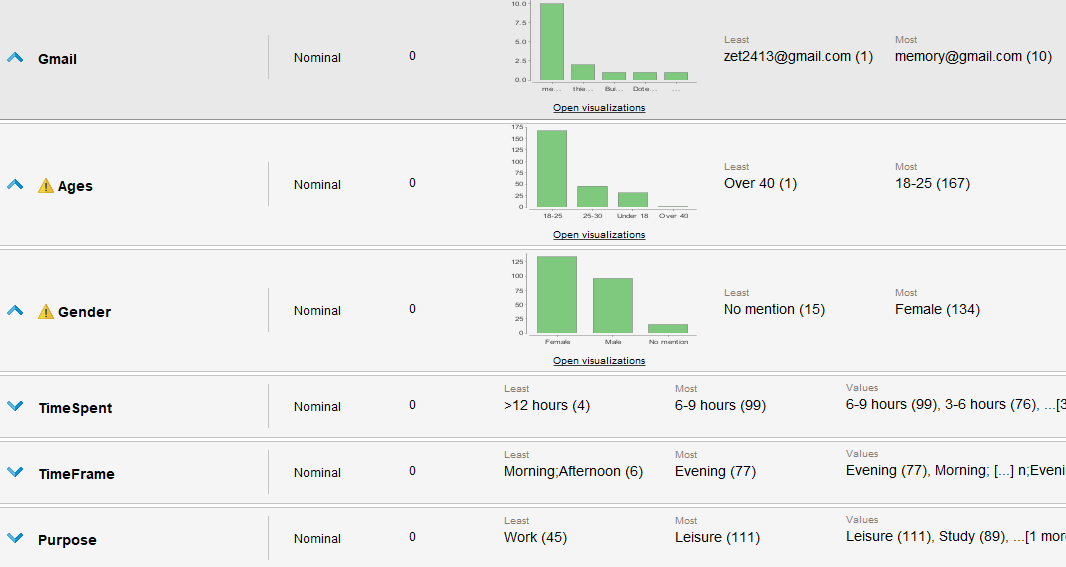
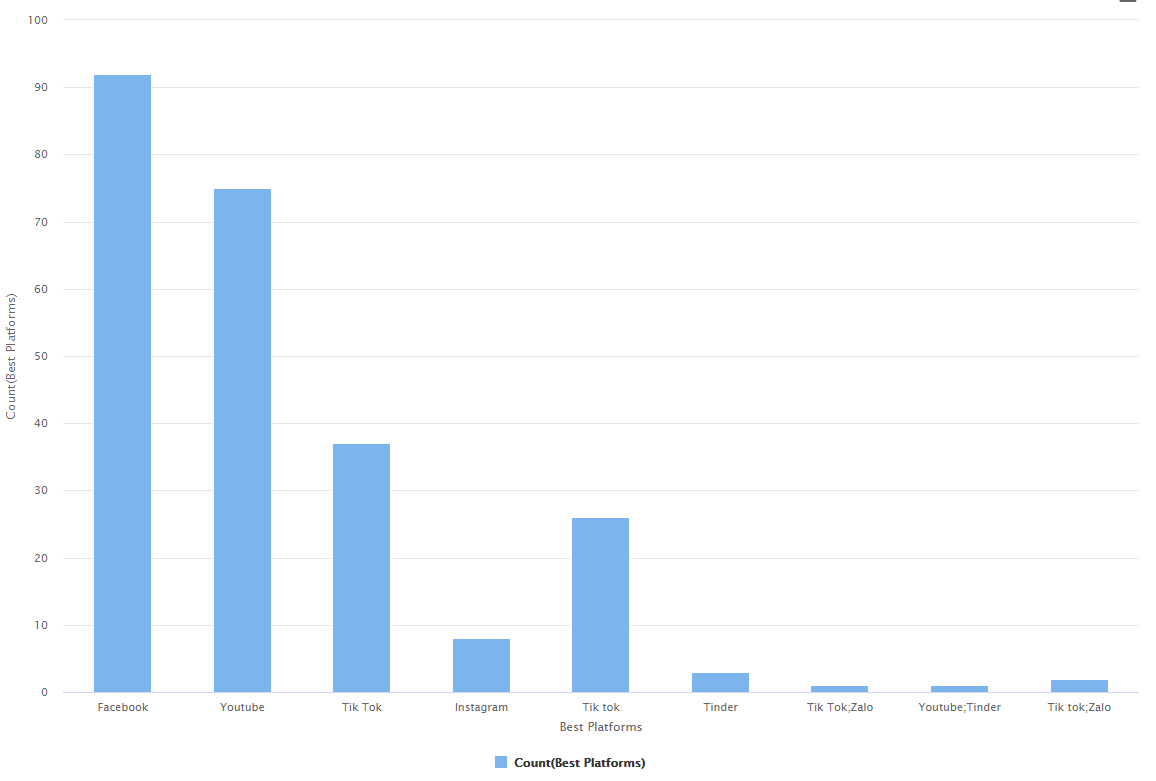
** **

Figure 3:Statistics Figure 4:Visualization

## b)Handle missing value

* When we load the raw data table into RAPIDMINER, we can plainly observe that many values are left blank and presented with a "?".

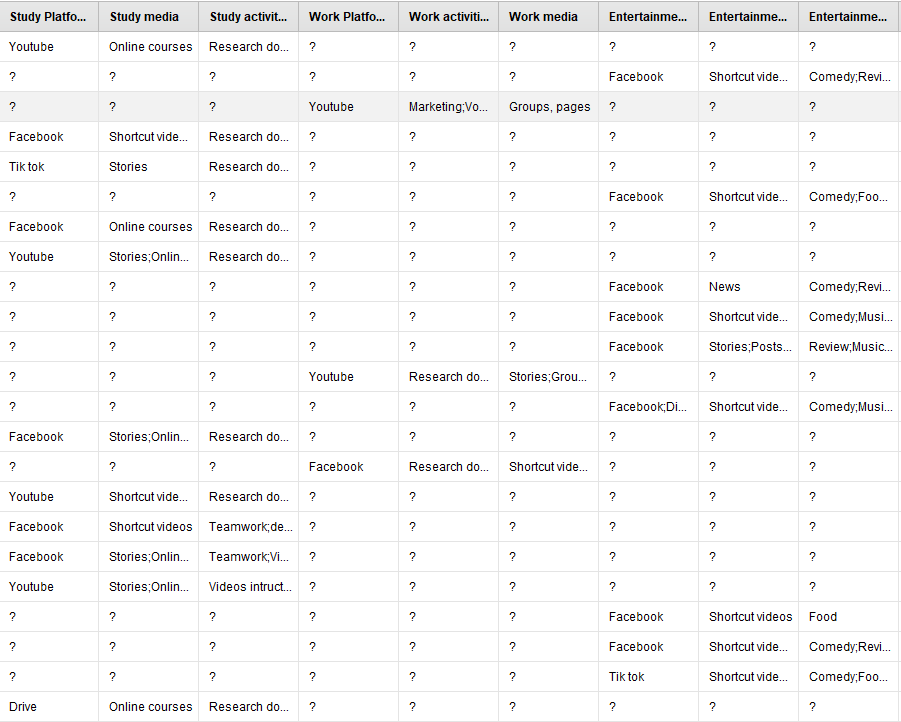


Figure 5: Missing value

* To remove the values "?" from the original *Rawdata*, we utilize the **Filter Examples** operations to remove them from the above data table**. Multiply** to obtain three distinct sorts of tables for three distinct purposes: leisure, work, and study.

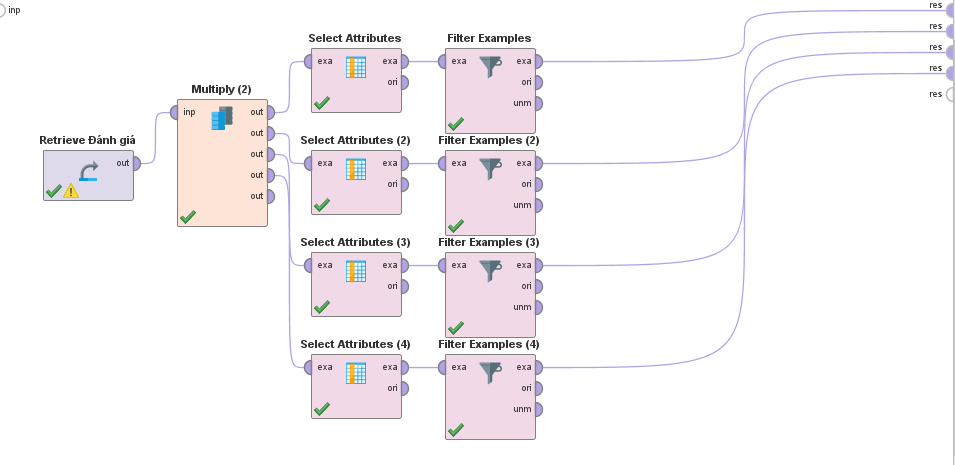
****

Figure 6 Rapidminer separation

* We offer three statistics tables for three different objectives and one table.

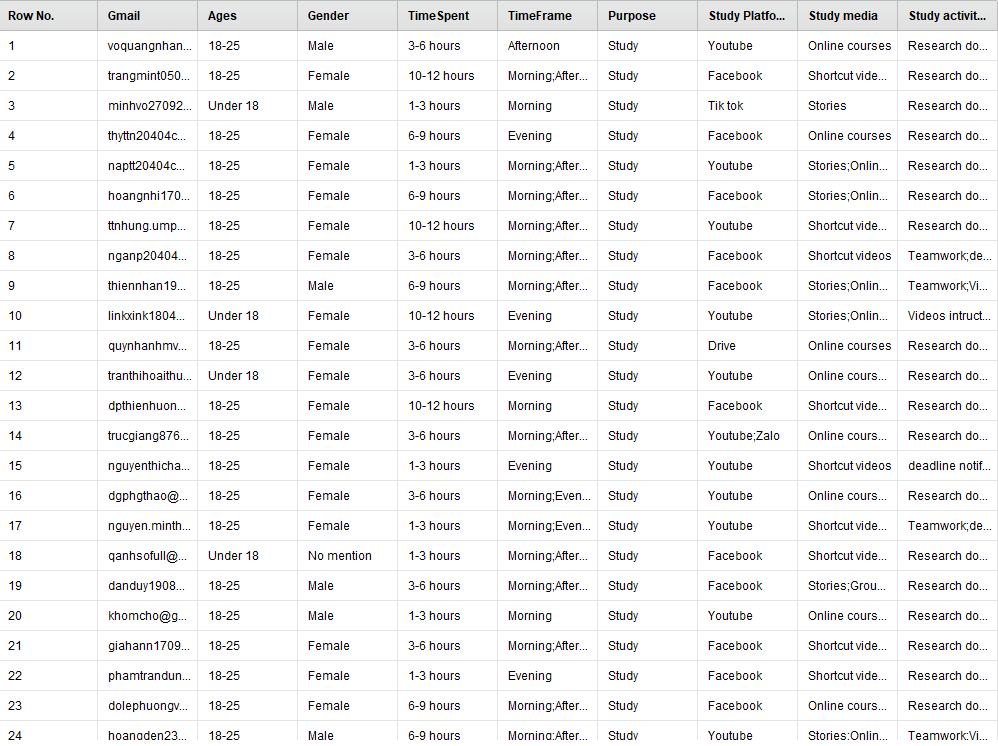
 

Figure 7: STUDY purpose Figure 8: WORK purpose

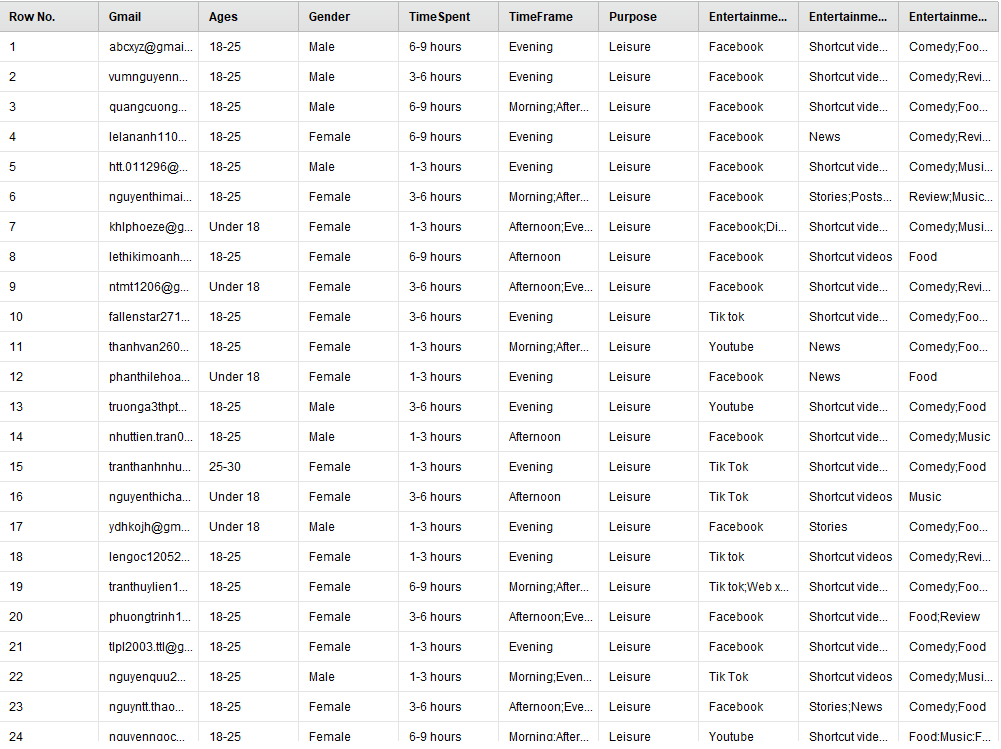


Figure 9: LEISURE purpose

* Besides, using tables for **Purpose**, adding two tables about **trustworthiness of social networking sites.**

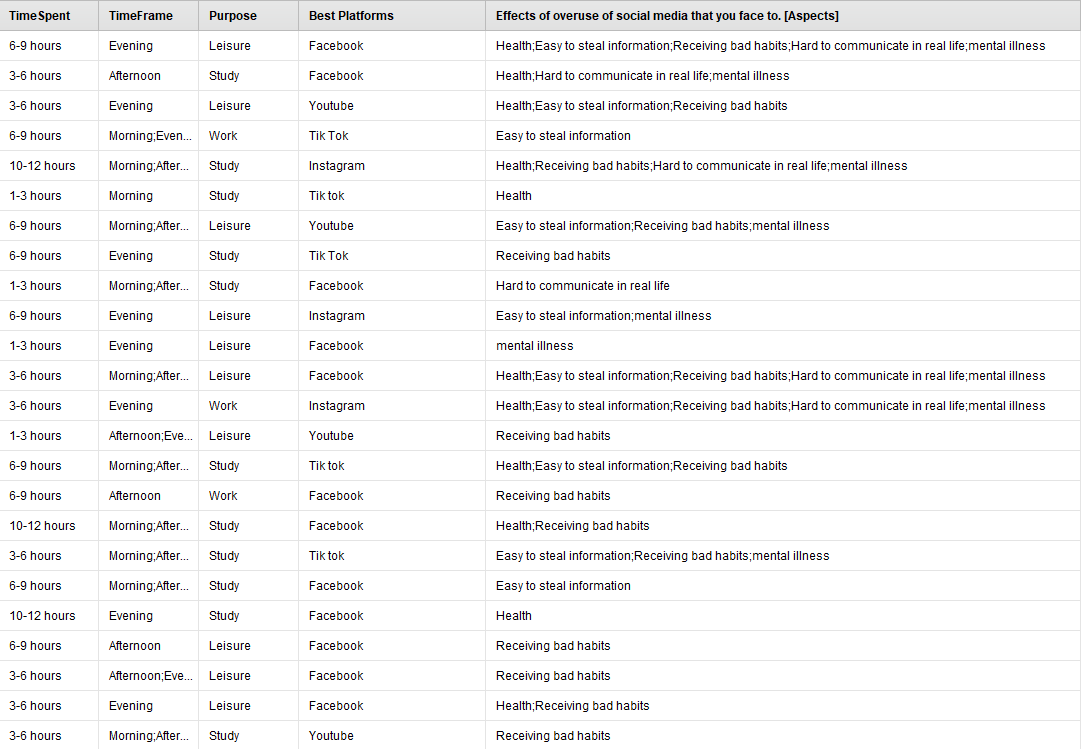
 

Figure 10:Effects Figure 11:Evaluation

## c)Build a predictive model

The sequence of the steps includes:

* + **Set Role**: I've set the Outcome property to Label, since I need the model to predict this outcome.
  + **Select Attributes:** drop the unnecessary columns
  + **Loop Operator:** Subprocesses are conducted five times, as specified by the number of citations parameter..
  + **Split validation Operator consists** consists of two distinct tasks: training (establishing a connection to the **R.F. model** using the default parameter) and testing (using **the Apply Model Operator** and **Performance**).

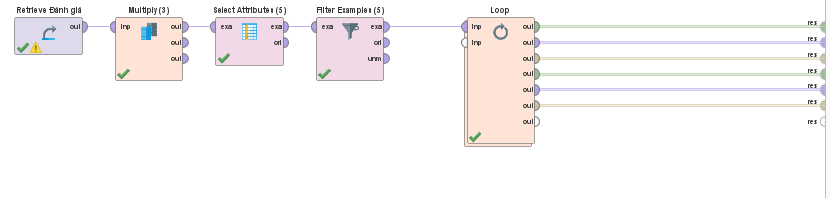


Figure 12: Using Loop Operator

* + **In Loop Operator consists** of 2 separate **Split validations**: One for Set Role **Ages label**, other is **Gender label**.

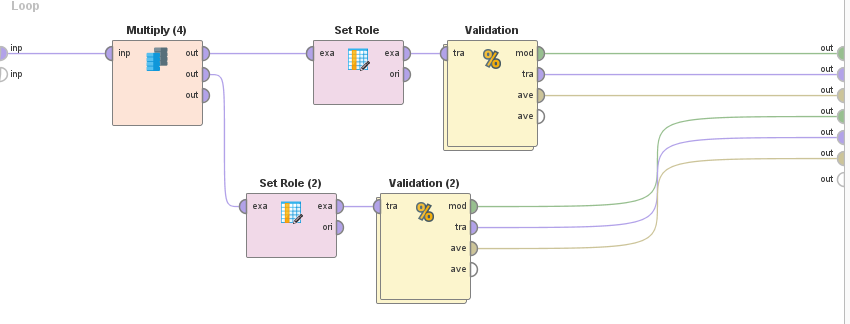


Figure 13: Using Split Validation

* We use Set Role to set labels in the branch that utilizes the Ages label, so using select attributes to remove superfluous columns such as just Ages, Best Flatforms, Gender, and Gmail.**l, Purpose, Timeframe, Timespent.**

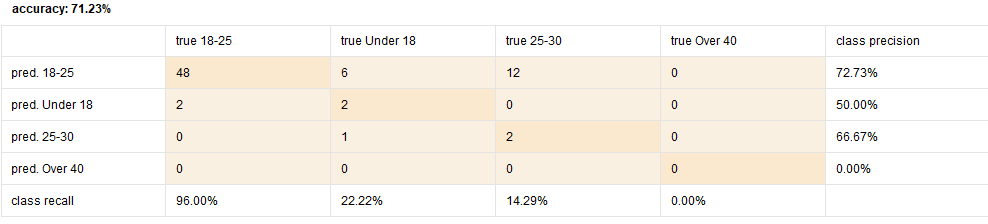


Figure 14: Age Lable Prediction

* The **Gender** label branch will utilize the same Operators as **the Set role**, but will modify the **Gender** labels.

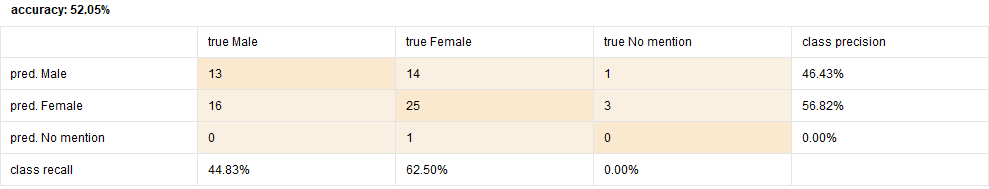


Figure 15: Gender label prediction

* Following that, in **Split validations**, we will employ two different kinds of **R.F models** on the Training side to predict two distinct labels for each branch, and two operators on the **Testing** side. **Applying the model** and **Performance** results in the generation of a set of performance criteria. the task classification's values.

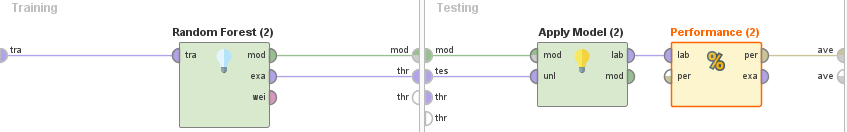


Figure 16: This Operator generates a decision tree model, which can be used for classification and regression.

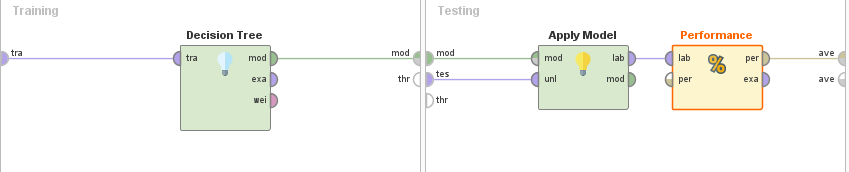


Figure 17: This Operator generates a random forest model, which can be used for classification and regression.

# IV. CONCLUSION

* We can discover relevant audiences for various types of content and promote to them accordingly using the qualities listed above.
* Regarding gender, it's rather straightforward to discern their primary reason for using social media, but it's more difficult to determine the particular they use (This section is not included in the information collection section). For instance, guys look for and utilize social media on a broader range of subjects than girls. Additionally, boys and girls may have the same personality and preferences whether it comes to a masculine or female theme.

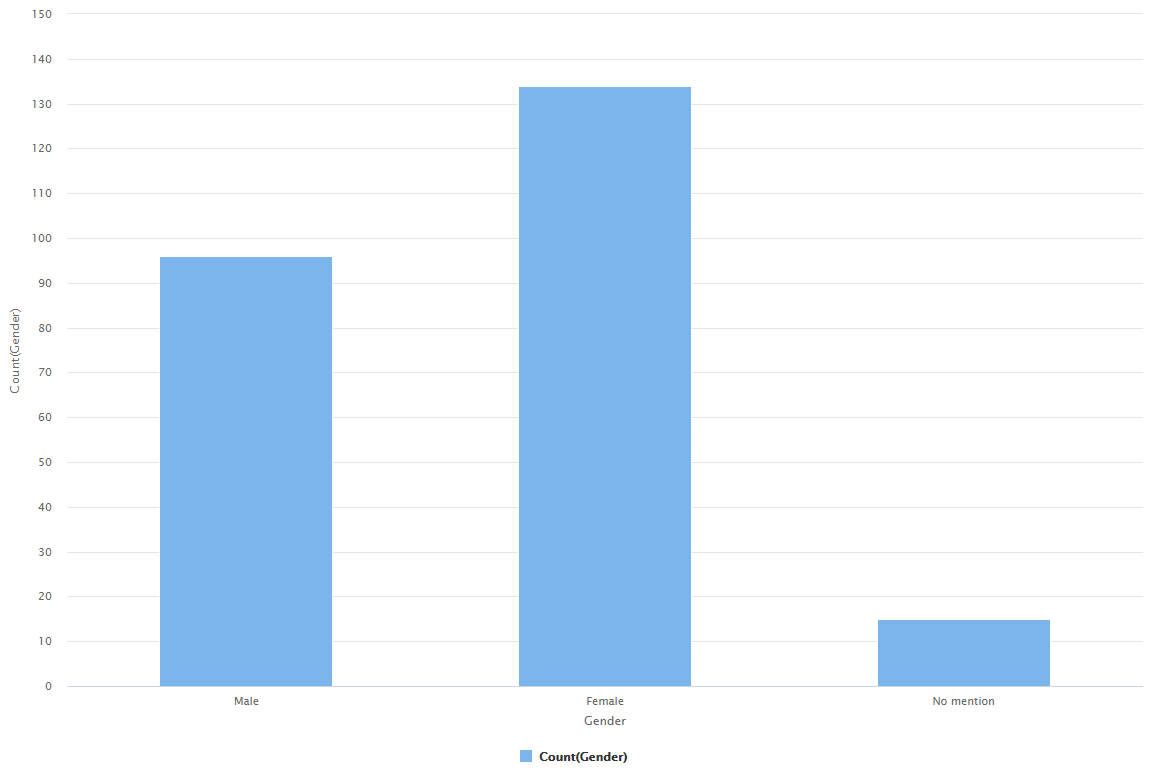


Figure 18: Gender Visualization

* The amount of time people spend on social media also helps us determine their intended use. Individuals who spend 6 to 9 hours on social media and 9 to 12 hours on social media have a variety of goals. They may study, work, or enjoy themselves, depending on how much time they spend on social media. Individuals who spend three to six hours on social media do so to study or work, sometimes connecting with friends or updating news. Individuals who spend between one and three hours prefer to focus on a single task, such as checking work and searching for job-related information, although they may also communicate with friends and view new things. Individuals who spend more than 12 hours a day on social media are extremely difficult to predict, since their time spent on social media is so extensive that their use is likewise diverse and complex. Those who worked 6 to more than 12 hours each day were often multi-taskers.

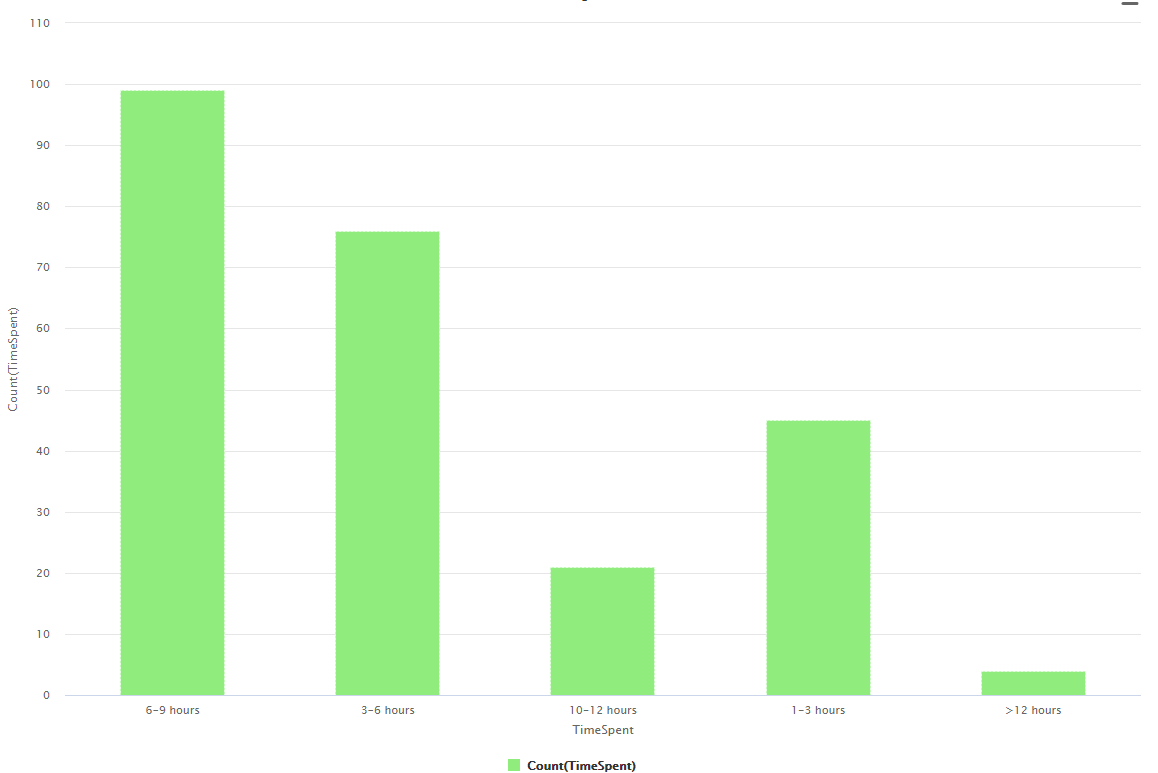


Figure 19 Timespent Visualization

* Between **the ages of 18 and 25**, people are either studying at university or doing a part-time job outside, therefore the primary aim of social media is to study and work**. Under-18s** are still in high school and haven't spent much time on social media; since then, the primary goal of this age group has been to amuse and converse with friends. Between **the ages of 25 and 30**, the majority of individuals work; their primary goal while using social media is to work; nevertheless, they may also gain soft skills or find new employment. courses relevant to their line of employment

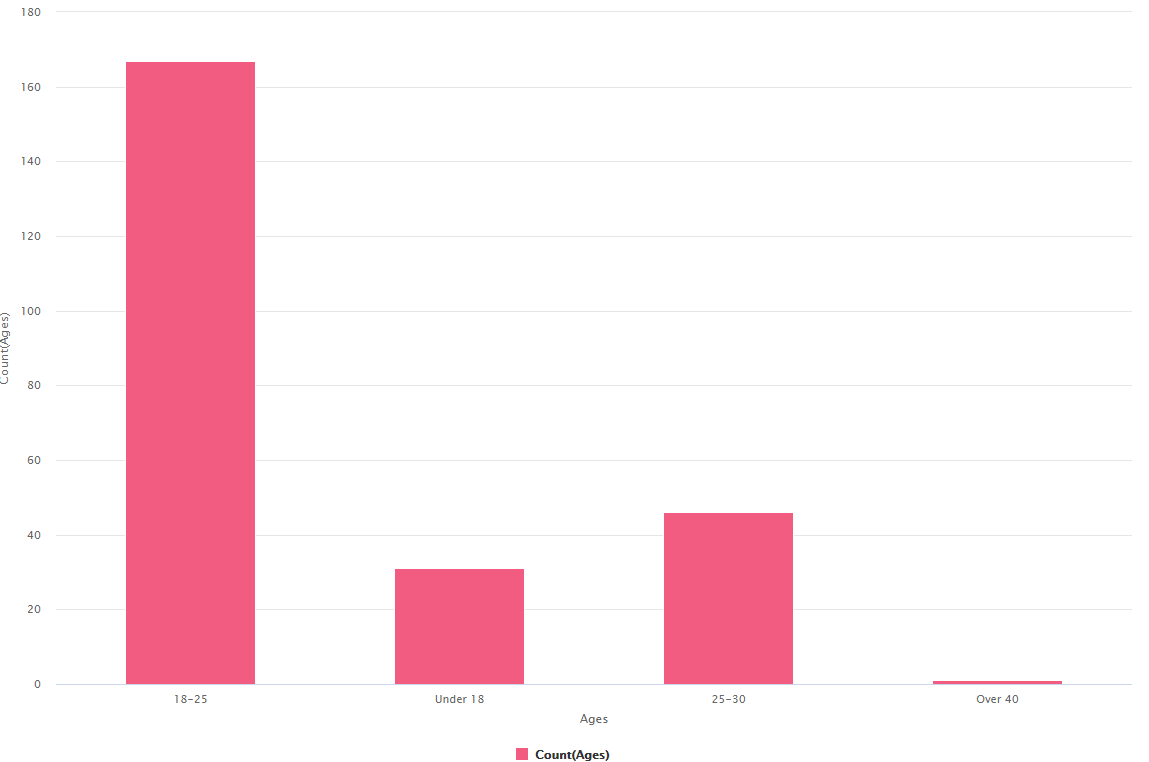


Figure 20: Ages Visualization

* People are most likely to use social media in the evening, after a long day at work and a desire to interact with friends or have fun. Apart from accessing social media in **the morning and afternoon**, they often work and study. They review the calendar of things that must be completed today, the topics, or do research on anything.

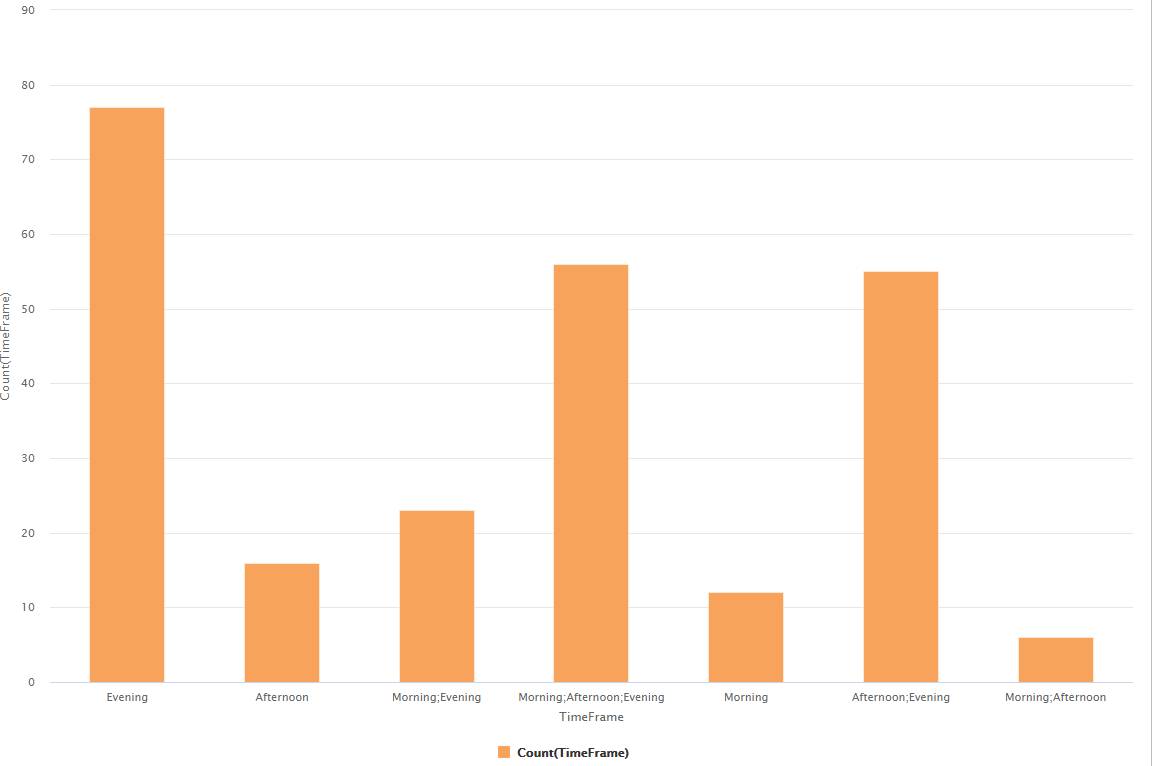


Figure 21: TimeFrame Visualization

## Difficulties and mistakes that may impair detection

* **Time:** There is no consistency in terms of time spent and time frame (e.g. 1-3 hours) when all three sessions are used (unrealistic and misleading in data.)
* The time spent between purposes is unaffected; people may use social media for a variety of reasons, not just one, such as pleasure, employment, or study, and therefore it is required to compare the portions hundred purposes together (which is missing in the above database).
* **Platforms:** The database contains information on the number of individuals who use it for specific purposes, not about a large social network.
* **Age**: Does not cover all age groups due to the target audience's lack of diversity.
* **Activities**: Because media may be accessed through search engines as well as social networks, selecting activities just via social media is not entirely genuine.
* The section on gathering user feedback on platform security informs us of the platform's degree of trust, from which we may deduce the platform's amount of utilization.
* **Most creative content**: Multiple alternatives complicate determining which social networks are the most innovative in terms of content.
* **Collecting data** using a **google form** results in an excel file that may have incorrect formatting, unequal information organization, and a large amount of missing data. As a result, data preparation will take a long time, which may easily result in data input mistakes. Additionally, acquiring raw data from social media is a significant challenge for us. The first is that the quantity of data available will be insufficient for me to accurately estimate the customer file in the future, and the error will be very significant.

## Predict

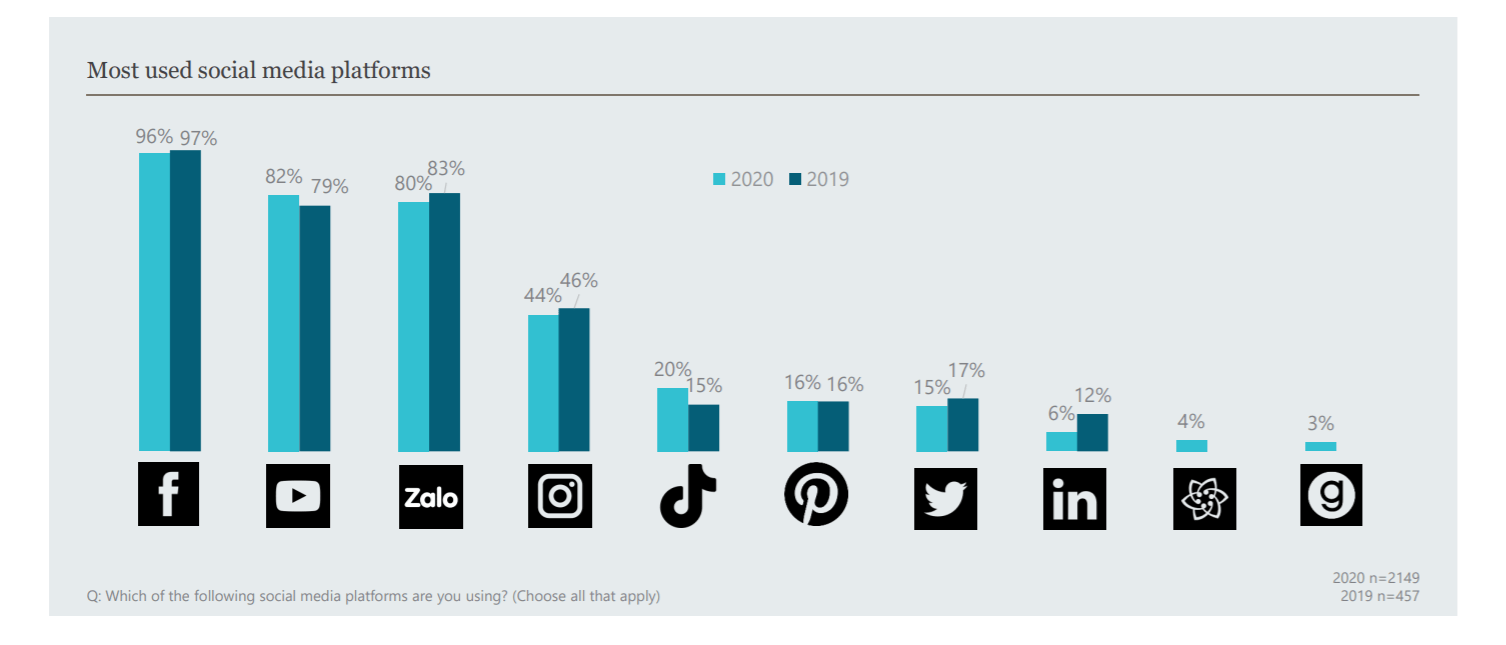
* + By using a predictive model of age and gender, it enables the media to target and select the most appropriate audience for their content and schedule. As a result, client trust is increased, and more effective advertising is used to offer items to a huge number of social network members. Since then, as a result of the customer's demands being addressed and trusted, the number of individuals utilizing social media has risen as well, as seen by:

Figure 22: Applied Platforms Visualization

[https://marketingai.vn/giai-ma-xu-huong-su-dung-mang-xa-hoi-cua- cac-the-he-nguoi-tieu-dung-viet-nam-trong-nua-dau-nam-2020/](https://marketingai.vn/giai-ma-xu-huong-su-dung-mang-xa-hoi-cua-%20%20%20%20%20%20%20%20%20%20%20%20%20cac-the-he-nguoi-tieu-dung-viet-nam-trong-nua-dau-nam-2020/)

* According to the factors predicted by Rapidminer, all of them result in very excellent results and will be much better when paired with an environmental survey.

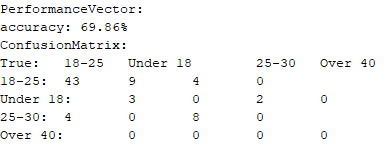


Figure 23: Data Prediction

* The assessment quality prediction and confidential method will aid in correctly evaluating social media at the moment and allocating the appropriate quantity of material to each social media platform. Since then, platforms with positive evaluations will be finished and reacted to progressively, and via that forecast, constrained social media platform will be addressed and conquered.

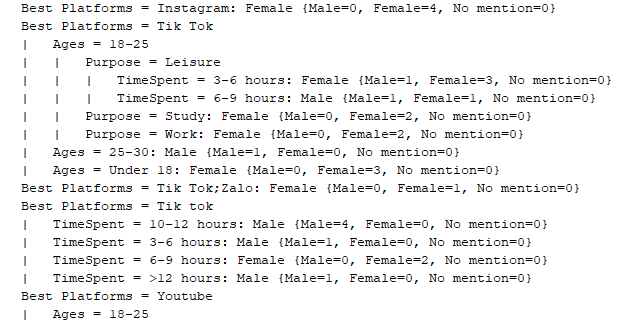
------------------------- 

Figure 24: Data Summary

* Additionally, this marketing method has a number of advantages, including increased open and click rates for your firm. This is because engaging clients through this communication channel enables you to engage in both brief and flashy advertising as well as intimate dialogues that have the ability to establish connections between you and your customers. As a result, it's probably worthwhile to experiment with a few different message platforms and advertising strategies to see which one works best for your organization.

# V.REFERENCES

* Creative advetising ideas:

<https://www.webcodebuddy.com/256-step-away-from-the-print-ad-creative-advertising-ideas-thatll-get-noticed/>

* Predict customer using Rapidminer Decision Tree:

<https://youtu.be/kRnjxgrg_PA>

* Predictive model Usages of social media meaning:

<https://marketingai.vn/giai-ma-xu-huong-su-dung-mang-xa-hoi-cua-cac-the-he-nguoi-tieu-dung-viet-nam-trong-nua-dau-nam-2020/>