Automatic Social Network Analysis System and Identification of Personality, Interests and Preferences of Users

Abstract

Social Networking Sites (SNS) affect people's daily lives, making them an important social stage for social interactions. Trust is an important aspect of relationships on social media. Whether it is used for security purposes, in identifying access to databases or in recommender systems, the definition of trust helps to develop algorithms for calculating trust relationships. In addition, a person's personality can be described as a set of standards that imposes a tendency on the individual's behaviour. Having information about a person's personality makes it easy to be guided in how they can react when faced with different situations. Identifying a user's personality can also facilitate knowledge of their potential needs in different situations. This work gathers data published on users' Facebook profiles, in a public group called Cheltenham Facebook Groups, proposes the creation of linear models, in order to calculate the level of trust between them, the creation of methods for predicting personality traits using the Big Five Personality model, as well as to identify their interests and preferences. Next, a Facebook scraper algorithm is introduced, which is used to extract users' public data from a Facebook API. The data collected is applied to the input of a classifier, called Ada Boosted Decision Tree, which uses a data mining technique to identify users' personality traits and interests from their profile, combining their public data with their answers to the Big Five Personality Test and Young People Survey respectively. The evaluation of the performance of the models, shows that in all cases satisfactory accuracy is achieved, which indicates with certainty, the ability of the models to measure and predict with high success rates, the level of trust between users and the analysis of their social behaviour in a network.

Keywords: machine learning, artificial intelligence, social networks, prediction models, linear regression

1. Introduction

In today's age of rapid technology and information, the Internet has emerged as the world's largest computer network. Facilitating the daily lives of billions of people, it is used to easily discover information and perform various tasks, making it quickly one of the main platforms for communication between people. The ability of the Internet to connect people in a short period of time has led to the creation of a large number of online social networks (OSNs), which consist of a collection of different types of social users, their relationships and interactions. The question that may arise is how the data posted by a user and his friends on online social networks can be used to assess the potential of their relationship, as well as further analysis of their social interactions, in order to determine of their personality and character, which can ideally facilitate the knowledge of their possible needs in different situations.

The purpose of this paper is to use the information posted by a user on his Facebook profile, in a public group called Cheltenham Facebook Groups, such as daily posts, reactions and messages, to suggest linear models in order to measure its level. his trust with his friends and the establishment of methods for measuring the score of his personality and interests, using, among others, the Big Five Personality Model.

In addition, a Facebook scraper algorithm is used to collect user data from a Facebook API. The data collected will then be used by a classifier, who uses a data mining technique to identify each user's personality traits and interests/preferences, resulting from the combination of his or her public data with his or her responses. through some specific electronic questionnaires.

The main questions that this research project tries to answer are: i) what does a general model for calculating trust between users look like globally, within a social network ii) how does this model differ from a model that calculates trust of a particular user with his closest friends iii) how these models can be applied using a particular social network as a

means (e.g. Facebook) iv) how approximate machine learning models can be developed to make predictions about personality and user interests through electronic surveys on the social network.

The personalized user data used in the survey forms a large data set, which focuses on the number of interactions and common social activities among Facebook users of Cheltenham Groups. By applying this data to the proposed models, conclusions can be drawn about their quality and accuracy, in order to determine if they can be used, so that they can safely predict the level of trust and relationship of users, but also categorize its characteristics. their personality and behavior, for a better understanding of their social behavior.

The present work is organized as follows: Section 2 describes the Relevant Tasks, Section 3 introduces the model for calculating the trust and the description of the research that will follow, Section 4 presents the results of the research and system analysis created, as well as the experimental evaluation of these results and Section 5 presents the conclusions produced by the research process.

2. Related Work

The trust measure shows the presence of a relationship between trustor and trustee. It is a subjective measure in which a given person A can have his own opinion of another person B, in order to consider the latter credible or not.

Yousra Asim, Ahmad Kamran Malik, Basit Raza και Ahmad Raza Shahid [1], in 2018, in the context of the trust calculation, they considered the direct and indirect trust of the user, which are based on the online interactions between the users. Their study proposed a SNTrust model to find the trust of users in a network, applying their proposed methodologies on online datasets and using methods (K-core, closeness centrality, eigenvector centrality, page rank), concluding that Their relationship of trust is linearly positive and statistically significant.

Moreover, in 2012, Vedran Podobnik, Darko Striga, Ana Jandras και Ignac Lovrek[2], proposed a new model for calculating trust in Facebook, which is verified through a Facebook application called "Closest Friends", which focuses on the similarity of the profile of an "ego" user with the profiles of his closest friends, taking into account the their joint activities. Through their research, they essentially aimed to suggest recommender systems for exploiting the potential of the social network by filtering information and offering suggestions to a user who is expected to like it.

In addition to defining the importance of close and intimate relationships between social media users, personality recognition is one of the new challenges among researchers on social media.

In 2018, Alireza Souri, Shafigheh Hosseinpour και Amir Masoud Rahmani, presented the hypothesis that users with similar personalities are expected to exhibit reciprocal patterns of behavior when collaborating through social networks [8], which was based on the answers given by users to a personality test (questionnaire NEO-FF-R-60), on their character and behavior. Based on all the data collected, they presented the Big Five Model, a personality model that classifies personalities with the help of classifiers, into five categories - Conscientiousness, Extroversion, Openness, Agreeableness and Neuroticism. And to identify these features, they used the AdaBoost method to enhance classification accuracy so that safe conclusions could be drawn for users.

In a similar way, a team consisting of Jennifer Golbeck, Cristina Robles and Karen Turner (2011) [10] and another of Michael M. Tadesse, Hongfei Lin, Bo Xu and Liang Yang (2018) [9], tried to understand how they can predict personality through social media profiles. In particular, they created applications on Facebook that use the Big Five Model just as effectively, but unlike other researchers, record users' personal information, focusing more on the features of their language, measuring the length of the character in each entry with language analysis tools, such as LIWC, SPLICE and SNA. Finally, with the application of the M5 'Rules and Gaussian Processes machine learning algorithms and the XGBoost classifier respectively, predictions were made for each personality trait.

3. Methodology

A social network can be considered mainly as a platform or website where different agents are encouraged to register, usually for free, and to connect with different agents participating in the same social network. This power of people to interact with other colleagues in an online environment is the main factor that contributes to the success or failure of such networks.

Facebook has been one of the most popular social networking platforms for years. In this network, many of the users (agents) may be physically unknown to each other or not. If two participants wish to communicate with each other, for various reasons, the assessment of their reliability, along a specific path of trust between them, within the social network

is mandatory. But the level of reliability can vary and is sometimes subjective and depends on the specific role of the individual in the network.

In this work, what is being attempted, whether easy or not, is to create models that define different levels of trust through mathematical formulas and algorithmic processes and suggest the basic idea behind the model for measuring personality traits.

3.1 Trust on a global level (GLOBAL TRUST)

Model for calculating global trust in a social network/group

Following the research of Yousra Asim, Ahmad Kamran Malik, Basit Raza and Ahmad Raza Shahid in 2018 ((Shahid, 2018)), in the same way it was possible to separate the process of calculating trust in direct (indirect) and indirect (indirect) contact with a user. In short, Direct trust shows the direct connection of a trustor with a trustee. In case they participate in common "activities", their relationship of trust can be measured. Similarly, but with some differences, Indirect trust between users can be calculated by measuring a user's activities, such as other users' responses / reactions to those activities.

The direct trust calculation has already been presented and evaluated in the paper of the previous group of researchers. Therefore, the goal will be how to approach and build a model, which will be able to effectively calculate the Indirect trust between users from the same Facebook group, dividing it first, into Participation and Response trust, as did the previous researchers in their research.

3.1.1 Participation Trust

Trust can be analyzed by the degree to which the user engages in network / group-level post sharing. If the user's total posts in a network / group are less than the average of the posts in that network / group respectively, the user is considered a Non-Participative, otherwise Participative. The Participation Trust (PT) of each user at network/group level is calculated by:

$$\mathrm{PT(u)} = \begin{cases} 1 \; if \;\; P(u) \geqslant AP \\ 0 \; otherwise \end{cases} \quad \textbf{(1)}$$

P(u): the number of posts of user u in network / group by: $P(u) = \sum_{i=1}^{n} Pi$ **(2)**

AP: the average number of posts per user at network / group level from: $AP = \frac{\sum_{i=1}^{n} P_N}{n_N}$ **(3)**

where the counter on the numerator indicates the total number of posts and n_N is the total number of users on the network.

3.1.2 Response Trust

Each user's trust is measured at the network / team level by examining their "reputation", which comes from the users who interact with it. If a user shares a post in a network / group, his Indirect Trust is calculated using the reactions he received in his posts from other member-users. For the Response Trust, the number of positive and negative responses (reactions) in the user's post is taken into account.

The reaction categories and their respective weights are listed in **Table 1**.

Table 1
User responses and weights on each post

POSSIBLE RESPONSES (acts)	RESPONSE WEIGHTS (w_{act})
SHARE	1
COMMENT	0.75
LIKE	0.25
LOVE	0.75
SAD	-0.50
ANGRY	-1
WOW	0.50
HAHA	0.25
THANKFUL	0.25

The Response Trust (ResT) of the user's post is calculated as the weighted sum of all reactions (positive or negative) to that post, divided by the total number of reactions to that post shown below:

$$ResT(u) = \forall\, v \in Network\left(\frac{\sum_{act \in ACT}(act(v_{up}) * w_{act})}{\sum_{\forall\, act \in ACT} up}\right) \quad \textbf{(4)}$$

where act (v_{up}) indicates the action (action) of user v in the post up of user u and w_{act} indicates the weight (weight) assigned to each action, through a set of actions. The average value of Response trust for a user u is calculated for all posts n in a network / group as follows:

$$avg.\ ResT(u) = \frac{\sum_{i=1}^{n} ResT_i(u)}{\sum_{i=1}^{n} up} \quad \textbf{(5)}$$

where ResTi (u) is u's Response Trust to a post. The sum of all user posts u is divided by the total number of posts u in network / group. If avg.ResT (u). 0.5 then u receives a good Response trust, otherwise not. The higher the value of the Response trust, it indicates that the user's posts are liked and considered worthy by other members of his team. The Indirect Trust (IT) of each user at the network / community level is calculated by combining Participation Trust (PT) and Response Trust (ResT).

The following equation shows the final calculation of the user's Indirect Trust, where $\omega = \mu = 0.5$.

$$IT(u) = \omega PT(u) + \mu.\ avg.\ ResT(u) \quad \textbf{(6)}$$

If IT (u) $\geq$ 0.5, user u is considered trusted at network/community level, otherwise not. Greater user trust indicates that the user is an active participant within the network/group and all user posts are valued by all other users. If a user does not join the network/group and does not receive feedback from other members of their group, then they are considered an inactive member.

3.2 Trust between pairs of friends (EGO TRUST)

Model for calculating ego trust in a social network

This section focuses on how social activities between an ego user and his friends can be used to calculate the "level of trust" between two friends at a time (= pair of friends) within a social network.

In this context, as defined by Vedran Podobnik, Darko Striga, Ana Jandras and Ignac Lovrek in 2012 in their research **(Vedran Podobnik, 2012)**, so here too a social activity is defined as a social interaction between two directly connected users (friends) in a social network. An example of social activity can be the joint tagging of two friends in the same photo. Now, we can define C ($user_x$) as a list of friends who commented on $user_x$ posts at a specific time period. Additionally, an element in C ($user_x$) is a pair:

($user_y$, number of interactions) **(7)**

where $user_y$ identifies a social network user who is directly connected to $user_x$ and who has had a number of interactions (= comments on posts) with $user_x$. For example, a C ($user_x$) list can be defined as:

$C(user_x) = \{ (user_a,3), (user_b,2), (user_c,5) \}$ **(8)**

which means that $user_a$ has commented 3 posts of $user_x$, $user_b$ in 2 posts and $user_c$ in 5 posts.

In addition, we define C ($user_x$, $user_y$) as the total number of comments in posts, between $user_x$ and $user_y$ friends within a social network.

In this example, we have $C(user_x , user_y) = 3$.

In order to calculate a "level of trust" between two directly connected users in a social network, it is necessary to create as complete a description of all social activities as possible between these friends.

Therefore, this point defines the descriptions of various social activities of the $user_x$ as the set of social activity ($user_x$), which is the set of lists where: i) each list describes a different social activity and ii) these lists are defined respectively as the list C ($user_x$).

In order to be able to calculate trust from interaction data contained in the social activity set ($user_x$), it is necessary to multiply the specific activities from each list in the social activity set ($user_x$) by specific weights for each type of interaction.

The following formula is used to calculate the level of trust between an ego user and each of his friends (the higher the level of trust, the stronger the relationship) on the same social network (for each $friend_x$ listed F ($user_x$)):

$$\text{trust (ego user, friend}_x) = \frac{\sum_{N \in social\ activity\ set(ego\ user)} W_N \times N(ego\ user, friend_x)}{\sum_{N \in social\ activity\ set(ego\ user)} W_N} \quad \textbf{(9)}$$

Thus, in order to use **Formula 9**, the first and most important part of the research is to collect and classify all relevant social activities from the ego user's Facebook profile.

Table 2 effectively classifies the activities that will be needed and their respective weights.

Table 2
$User_x$'s friend lists sorted by social activities on the Facebook social network

List Label	Description of the list	Weight (W_N)
C($user_x$)	List of friends who comment on $user_x$ ´s Facebook posts	0.75
CL($user_x$)	List of friends who react with 'LIKE' on $user_x$ ´s comments	0.25
LIKE($user_x$)	List of friends who react with 'LIKE' on $user_x$ ´s posts	0.25
LOVE($user_x$)	List of friends who react with 'LOVE' on $user_x$ ´s posts	0.75
SAD($user_x$)	List of friends who react with 'SAD' on $user_x$ ´s posts	-0.50
ANGRY($user_x$)	List of friends who react with 'ANGRY' on $user_x$ ´s posts	-1
WOW($user_x$)	List of friends who react with 'WOW' on $user_x$ ´s posts	0.50
HAHA($user_x$)	List of friends who react with 'HAHA' on $user_x$ ´s posts	0.25
THANK($user_x$)	List of friends who react with 'THANKFUL' on $user_x$ ´s posts	0.25
F($user_x$)	List of all friends of the Facebook $user_x$	---

All the lists from **Table 2** (except list F ($user_x$)) are a complete set of social activities between $user_x$ and all his Facebook friends:

social activity set($user_x$) = **{C**($user_x$), **CL**($user_x$), **LIKE**($user_x$), **LOVE**($user_x$), **SAD**($user_x$), **ANGRY**($user_x$) , **WOW**($user_x$) , **HAHA**($user_x$) , **THANK**($user_x$)**}** **(10)**

Each list from the above set of social activities ($user_x$) indicates a number of pairs ($user_x$, $user_y$), which is described in **Formula 7**.

What **Table 2** shows with simple examples is that if there are 20 different friends who commented on the same post on $user_x$'s Facebook profile, the **C** (ego user) list will consist of 50 pairs, as reported in Formula 7. Similarly, if 5 different friends react with "LIKE" to a post that $user_x$ posted on his Facebook wall, then the **LIKE** (ego user) list will consist of 5 pairs, but if there are no friends who reacted with "ANGRY" »In a $user_x$ post, the **ANGRY** (ego user) list will be empty, which means that ANGRY (ego user) = 0.

3.3 Big-Five Personality Model

Five Factor Model (FFM)

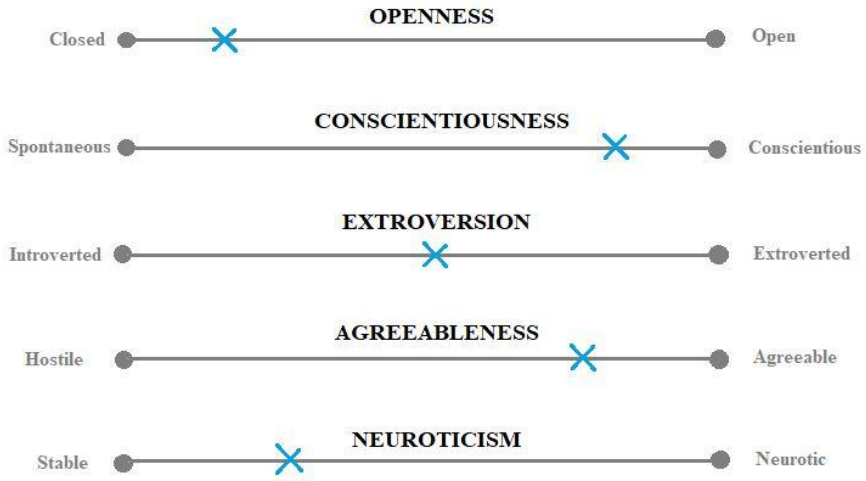
According to Criteriacorp.com[1], ‹‹ Personality research has created a variety of different theories that try to define and measure personality. The most widely accepted classification of personality among industrial-organizational psychologists is the Big Five Personalities model, or the Five Factor model of personality. The Five Factor model divides personality into five elements: Agreeableness, Conscientiousness, Extraversion, Openness and Neuroticism. Personality tests based on this model measure the range of each of the five characteristics of each person. ››

Each of the characteristics measures a unique aspect of the human personality:

[1] https://www.criteriacorp.com/resources/glossary/big-five-personality-traits

- **Extraversion** is a measure of how social, extroverted and energetic a person is. People who have a lower score on the extroversion scale are considered to be more introverted, quiet and low-key.
- **Agreeableness** is a measure that shows an individual's relationship to social harmony. It basically shows how well the person gets along with others, how cooperative they are and how they can interact within a group.
- **Conscientiousness** is a measure of how careful, disciplined, and organized a person is.
- **Neuroticism** is a measure that shows the ways in which people react emotionally to different situations.
- **Openness to Experience** is a measure of how imaginative and creative a person is, or how down to earth and conventional they can be.

‹‹ Contrary to other theories about personality traits that classify individuals into binary categories, the Big Five Model claims that each personality trait is a spectrum. Therefore, individuals are classified on a scale between the two extremes. ››[2]



A simple example illustrating the results of the model
(image source: https://www.simplypsychology.org/big-five-personality.html**)**

3.3.1 Application of the Big-Five Personality Model

In order to accurately measure the social behavior of Facebook users, it was necessary to gather the necessary data, with the answers given on the Open-Source Psychometrics Project[3], an online questionnaire that includes the Big Five model.

The term Big Five comes from the statistical study of responses to personality traits. Using a technique called factor analysis, researchers can look at people's responses to hundreds of personality items and ask the question, "What is the best way to summarize a person?" This has been done with many samples from around the world and the general result is that, while there seem to be unlimited personality variables, five stand out from the pack in explaining the individual's many answers to questions about his personality: extroversion, neuroticism, agreeableness, conscientiousness and openness to experience. The Big Five is not related to any specific test, but instead a variety of measures have been developed to measure it. This particular test uses the Big-Five Factor Markers from the International Personality Item Pool, developed by Goldberg (1992).

In short, the Big Five Personality Test consists of 50 items (as shown in **Table 3**) that each user must evaluate how true it is for him, on a five-point scale where 1 = Disagree, 3 = Neutral and 5 = Agree.

[2] https://www.simplypsychology.org/big-five-personality.html
[3] https://openpsychometrics.org/tests/IPIP-BFFM/

After the user answers the questions, his total score for each feature based on the positive or negative value of each question, is calculated as described below:

Extraversion = 20 + EXT1 – EXT 2 + EXT 3 – EXT 4 + EXT5 – EXT6 + EXT7 – EXT8 + EXT9 – EXT10

Agreeableness = 14 – AGR1 + AGR2 – AGR3 + AGR4 – AGR5 + AGR6 – AGR7 + AGR8 + AGR9 + AGR10

Conscientiousness = 14 + CSN1 – CSN2 + CSN3 – CSN4 + CSN5 – CSN6 + CSN7 – CSN8 + CSN9 + CSN10

Neuroticism = 38 – N1 + N2 – N3 + N4 – N5 – N6 – N7 – N8 – N9 – N10

Openness = 8 + OPN1 – OPN2 + OPN3 – OPN4 + OPN5 – OPN6 + OPN7 + OPN8 + OPN9 + OPN10

Table 3
The 50 Personality Test questions categorized by attribute

Extraversion (EXT)	**Agreeableness (AGR)**	**Conscientiousness (CSN)**	**Neuroticism (EST)**	**Openness (OPN)**
I am the life of the party.	I feel little concern for others.	I am always prepared.	I get stressed out easily.	I have a rich vocabulary.
I don't talk a lot.	I am interested in people.	I leave my belongings around.	I am relaxed most of the time.	I have difficulty understanding abstract ideas.
I feel comfortable around people.	I insult people.	I pay attention to details.	I worry about things.	I have a vivid imagination.
I keep in the background.	I sympathize with others' feelings.	I make a mess of things.	I seldom feel blue.	I am not interested in abstract ideas.
I start conversations.	I am not interested in other people's problems.	I get chores done right away.	I am easily disturbed.	I have excellent ideas.
I have little to say.	I have a soft heart.	I often forget to put things back in their proper place.	I get upset easily.	I do not have a good imagination.
I talk to a lot of different people at parties.	I am not really interested in others.	I like order.	I change my mood a lot.	I am quick to understand things.
I don't like to draw attention to myself.	I take time out for others.	I shirk my duties.	I have frequent mood swings.	I use difficult words.
I don't mind being the center of attention.	I feel others' emotions.	I follow a schedule.	I get irritated easily.	I spend time reflecting on things.
I am quiet around strangers.	I make people feel at ease.	I am exacting in my work.	I often feel blue.	I am full of ideas.

And lastly, as **Table 4** depicts, if the user's total feature score is higher than the average feature score in the network, then his personality result is assigned with 1 and if it is lower, it is assigned with 0.

Table 4
The interpretation behind the final personality score

Personality Factors	**Result of Personality Factor = 0**	**Result of Personality Factor = 1**
Extraversion	Dissociable, Solitary	Sociable, Warm
Openness to Experience	Strong, Tenacious	Conventional
Agreeableness	Negative	Accessible, Receptive
Neuroticism	Nervous	Calm

Conscientiousness	Selfish	Helpful

3.4 Online Research of Interests

In addition to the user personality analysis previously presented in section 3.3, users' interest profiles can be analyzed and predicted using knowledge and data from their profiles that they have expressed interest in.

To this end, it was once again necessary to gather the necessary data, with the answers given to questions of an online questionnaire, in order to better clarify the interests and preferences, i.e., to make an easier understanding of which are the preferences of users in music, movies, their daily activities and hobbies.

In short, this online survey consists of a number of questions in specific categories (as shown in **Table 5**), which each user has to rate for himself, on a five-point scale where 1.2 = Disagree, 3 = Neutral and 4.5 = I agree. In case the score of a question after the end of the survey is blank, it means that the user has chosen not to answer the specific question and thus the score is automatically filled with 0.

Table 5
The main interest categories of the questionnaire

Interests	Questionnaire
Music	I enjoy listening to music
Movies	I really enjoy watching movies
Musical Instruments	I am interested in playing musical instruments
Fun with Friends	I enjoy spending time with my friends
Internet	I am interested in using the Internet
Countryside/Outdoors	I enjoy being at the countryside
Sports	I am interested in participating in sports activities

4. Experimental Study

In Section 3, the necessary formulas and methodologies were presented, which will be used to measure the level of trust between users (friends) from the same Facebook community, both globally and more directly.

In this section, the analysis of the proposed trust models is performed in a set of social network data available on the Internet, the Cheltenham Facebook Groups.

Contains aggregate data from 5 public groups, named Unofficial Cheltenham Township (Group1), Elkins Park Happenings (Group2), Free Speech Zone (Group3), Cheltenham Lateral Solutions (Group4) and Cheltenham Township Resident (Group5). This data set consists of data for each of these groups regarding group-level posts, comments on each post, likes and responses to posts, and group member information.

4.1 Assessment of trust models

4.1.1 Linear Regression

In the field of statistics and machine learning, Linear Regression[4] defines as a linear approach for modeling the relationship between a graded response of one or more explanatory variables, known as dependent and independent variables. The case of an explanatory variable is called ***Simple Linear Regression*** and for more than one, the process is called ***Multiple Linear Regression***.

[4] https://en.wikipedia.org/wiki/Linear_regression#Generalized_linear_models

In general, in linear regression, relationships are formed using linear predictive functions, the unknown parameters of which are estimated from the data provided. Such models are called linear models.

In the research conducted, the goal for using these models is first to find the desired relationship between the data variables and then to use this relationship to predict the outcome of a future "event".

4.1.1.1 Simple Linear Regression

According to the *Department of Statistics and Data Science* from Yale University, ‹‹ The simplest case of regression, which includes a single graded response variable prediction x and a single gradient response variable y is known as a simple linear regression. The basic model of simple linear regression follows an equation of the form $\boldsymbol{Y = a + bX}$, where **X** is the independent (explanatory) variable and **Y** is the dependent variable. The slope (= slope) of the line is b and a is the intersection (= intercept), i.e., the value of y when x = 0. ››[5]

4.1.1.2 Multiple Linear Regression

Respectively, according to the research from the *Department of Statistics and Data Science* , ‹‹The generalization of single linear regression in the case of more than one independent variable x and a special case of general linear models, limited to a dependent variable y, is known as multiple linear regression. The basic model of multiple linear regression follows an equation of the form $\boldsymbol{Y_I = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_I}$ for I = 1,2, ... n, where $\boldsymbol{X_I}$ are the independent variables and YI is the dependent variable for each i. ›› [6]

4.1.2 Methods of Linear Regression Models

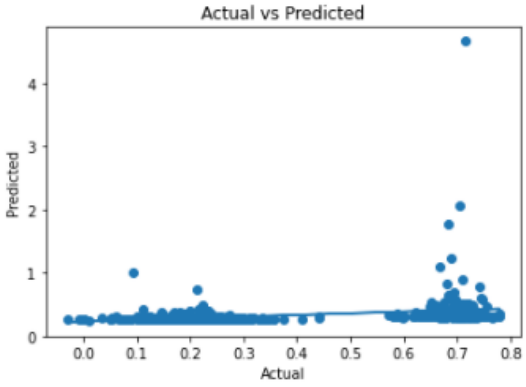
As already stated, a linear regression model generally attempts to model the relationship between variables by adapting a linear equation to certain data.
Before starting the process of adapting the linear model to the data, it must first be determined whether there is a relationship between the variables.
The following methods were used to construct the trust models mentioned above in the work in paragraphs 3.1 and 3.2, using data from the Cheltenham dataset:

i) Scaling is a method used to "normalize" the range of independent variables or data characteristics, while also being called data pre-processing. This method is a good way to find the data in the same state, assigning it to new values, so that it is easier to compare with each other.
ii) ***Train / Test*** is a method of measuring the accuracy of a model. It is called Train / Test because the data is divided into two sets: a training set (Train) and a set of tests (Test). In this research, the dataset was divided into 70% for training in order to create the models and 30% to test their accuracy.
iii) ***R-squared (r2_score)*** is a statistical measure that represents the good fit of the regression model. The ideal value for r-squared is 1. The closer its value is to 1, the better the model.
iv) A ***scatter plot*** diagram is a diagram where each value in the dataset is represented by a dot / dot. It is a useful tool that helps to determine the strength of the relationship between variables. If in the end there does not seem to be a correlation between the proposed independent and dependent variables, then the adaptation of the linear regression model to the data probably does not offer a useful model.
Applying the formulas of Modules 3.1 and 3.2 to the different variants of Cheltenham Facebook Groups data, depending on the type of information available, we try to determine exactly which of the Indirect (A, B, C) and Ego (D, E) models Trust respectively are the most appropriate and then which of these products will eventually be used for the most efficient forecast of trust, always according to the needs of our research.

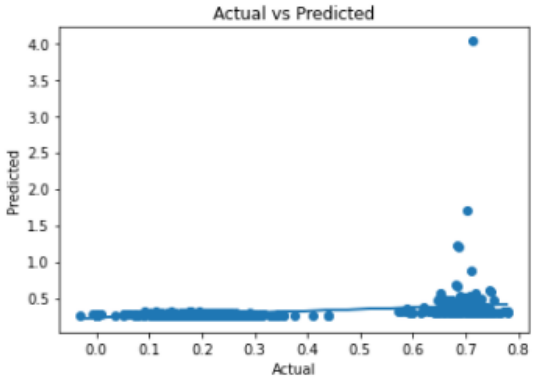
[5] http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm
[6] http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm

From the data generated by the mathematical formulas of these modules and which have been successfully pre-processed, divided into Train / Test and the dots shown in the corresponding diagram, it is easy to draw the lines of multiple linear regression, representing the different variations of models we consider.

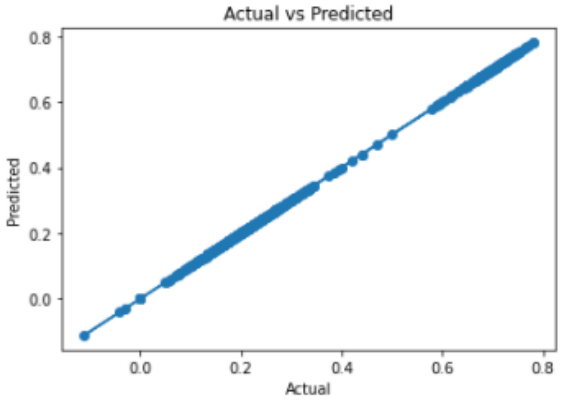
4.1.2.1 Indirect Trust Model



A)



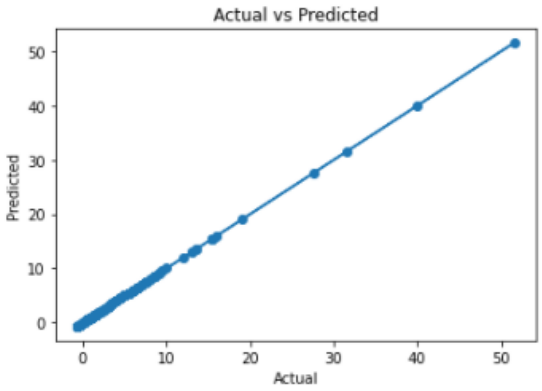
B)



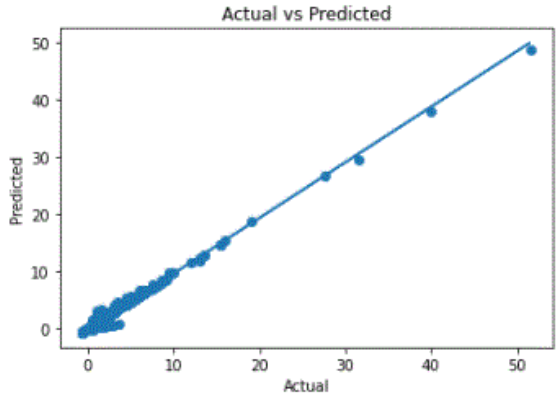
C)

GRAPH	DATA INPUTS	DATA OUTPUT	R2_SCORE
A	user_posts, user_comments, user_shares, user_likes, user_sad, user_love, user_angry, user_wow, user_haha, user_thankful	trust	0.234561
B	user_posts, total_responses	trust	-0.040403
C	user_posts, Participation Trust, Response Trust	trust	0.999999

4.1.2.2 Ego Trust Model



D)



E)

GRAPH	DATA INPUTS	DATA OUTPUT	R2_SCORE
D	total_responses_weight	trust	0.999999
E	comments, comment_likes, like,love,sad,angry,wow,haha,thankful	trust	0.984060

After creating models for calculating trust on a global and more interpersonal level, but also for creating and evaluating appropriate trust prediction models for new users in Facebook groups, the conclusion that emerges is that the most appropriate trust prediction model , which better meets the needs of the final system and on which the other models will be based later, is what calculates the Indirect Trust (**case C**), since it contains more important information about the profile of each user and in a very high accuracy, of 99%.

This is because we can now draw conclusions about whether a user is actively involved in his community, about whether the posts he uploads to his Facebook profile are accepted and liked by his friends, which ultimately shows whether he is trustworthy or not. from the rest of his fellow human beings.

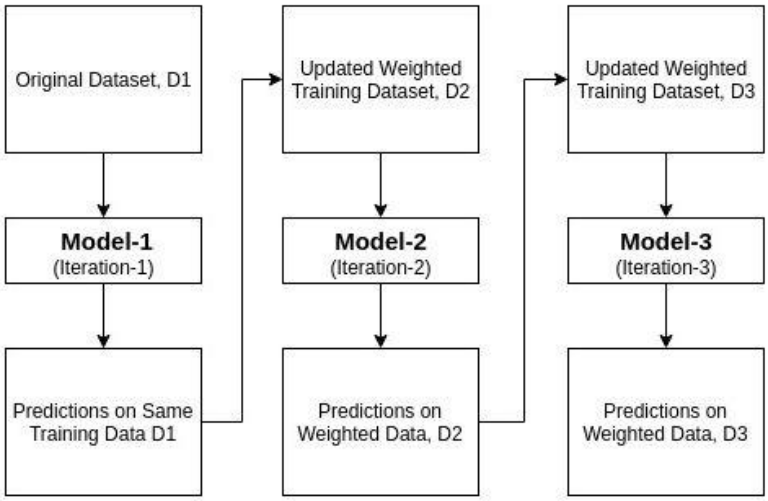
4.2 ADABOOST CLASSIFIER

According to the results of the 2018 survey, by Alireza Souri, Shafigheh Hosseinpour and Amir Masoud Rahmani **(Alireza Souri, 2018)**, the Ada Boost- Decision Tree model was the most accurate of the models they studied, with an accuracy of 82.2%, which aimed to predict the personality of users according to the variables in their profile. [7]
Following the example of their research, this work is based on the AdaBoost classifier for the identification of personality traits based on the 5 factors of the Big Five Personality test.
According to Yoa Freund and Robert Schapire, who won the Gödel Prize (2003) for their work, ‹‹***AdaBoost***, short for Adaptive Boosting, is a statistical classification algorithm. It can be used in conjunction with many other types of learning algorithms with the primary purpose of improving model performance. In essence, it uses low-precision classifiers in order to create a new high-precision classifier, that is, with the lowest possible error rate.
The AdaBoost algorithm involves the use of very short (level) decision trees that are added sequentially to the set. Each subsequent model tries to correct the predictions made by the model before. This is achieved by weighting the training data set to place more emphasis on training examples in which previous models made prediction errors. ›› [8]

[7] https://link.springer.com/article/10.1186/s13673-018-0147-4

How does the AdaBoost algorithm works internally? [9]

1. First, AdaBoost selects a random training subset.
2. Repeats the AdaBoost machine learning model by selecting the training set based on the exact prediction of the latest training process.
3. Assigns the highest weight to incorrectly sorted observations, so that in the next iteration these observations have a high probability of classification.
4. Thus, it assigns the weight to the trained classifier in each iteration depending on the accuracy of the classifier. The most accurate classifier will gain high weight.
5. This process is repeated until the complete training data match without any errors or until the specified maximum number of assessors is reached.
6. For the classification process, you just need to run a "vote" on all the learning algorithms created.



Graph showing the operation of the algorithm
(image source: https://python.plainenglish.io/adaboost-classifier-in-python-8d34a9f20459 **)**

4.3 Application of the classifier on the Cheltenham dataset

4.3.1 Personality Prediction

The personality prediction for each of the 2538 members of the Cheltenham teams can be measured by applying specific data as an introduction to the Ada Boosted Decision Tree algorithm. More specifically, from each data file, the most necessary for research and those that most effectively predict personality traits are: user_posts, user_comments, user_shares, user_likes and trust variables as X inputs, which show the number of posts the user has uploaded in his profile, the number of comments, shares and "LIKE" markings on his posts respectively and the SCORE variable, which has a value = 0 or 1 (the results are shown in **Table 4**), as Y output.
After training and testing the data for each personality trait and applying the AdaBoost Classifier, it is very important to determine - *How often is the classifier correct ??* - calculating the accuracy of the model.

Table 6
Predictive accuracy for each personality trait

Personality Trait	Accuracy

[8] https://en.wikipedia.org/wiki/AdaBoost
[9] https://python.plainenglish.io/adaboost-classifier-in-python-8d34a9f20459

Extraversion	~1.0
Agreeableness	~1.0
Conscientiousness	~0.99
Neuroticism	~0.99
Openness to Experience	~0.99

By collecting the results of the model for all the personality scores of the users in the group, a percentage of how many of them have "positive" and "negative" social behavior.

4.3.2 Interests Prediction

Accordingly, the prediction of interests for each of the 1010 members of the Cheltenham teams can be measured by applying specific data as an introduction to the Ada Boosted Decision Tree algorithm.
More specifically, from the data file, the most essential elements for research and those that best predict the interests and preferences of each user are: user_posts, user_comments, user_shares, user_likes and trust variables as X inputs, which show the number of the posts that the user uploaded to his profile, the number of comments, shares and "LIKE" tags of his posts respectively and the variables: Music, Movies, Musical instruments, Fun with friends, Internet, Countryside / Outdoors, Adrenaline sports , which have values from 1-5 (rarely maybe 0), as Y output.
After training and testing the data for each category and applying the AdaBoost Classifier, it is very important to determine - *How often is the classifier correct ??* – calculating the accuracy of the model.

Table 7
Accuracy of forecast for each interest category

Interests	Accuracy
Music	~ 0.84
Movies	~ 0.73
Musical Instruments	~ 0.50
Fun with Friends	~ 0.70
Internet	~ 0.60
Countryside/Outdoors	~ 0.72
Sports	~ 0.82

By collecting the results of the model for all user interest scores in the group, a percentage of how many of them have rated the questions on a scale of 0-5, as well as what this rating might mean for them.

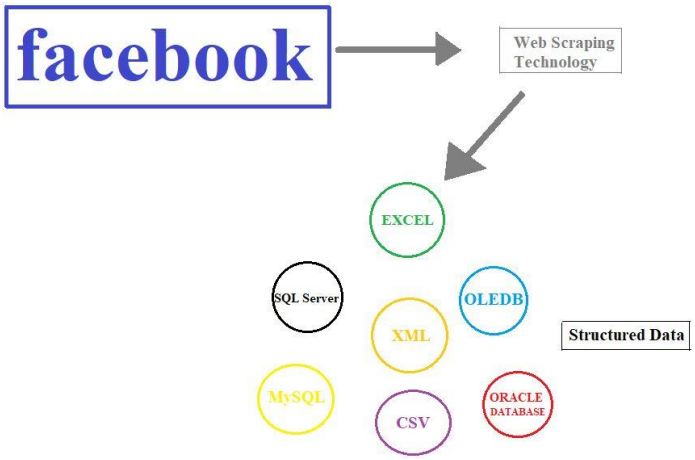
4.4 Application of the classifier on user data

After defining the methods and algorithms to be used to create and train the necessary standard for measuring and predicting the personality traits of the given user in Sections 3 and 4, it is important to evaluate its efficiency and accuracy. model performing all the procedures, with the only difference that now the collected data is imported from the user's Facebook profile. This can be achieved through data crawling or scraping processes.

As the editors on *promptcloud.com* dictate, ‹‹ *Data crawling* means working with large data sets, where one can develop crawlers (or bots) that crawl to the "deeper parts" of web pages. The *data scraping* process, on the other hand, refers to retrieving information from any source (not necessarily on the Web). ››

4.4.1 Facebook Data Scraping

Facebook is a huge database of user-generated content. Facebook data can be used by other researchers to better understand user behavior. This kind of research is what the Cambridge Analytica [10] introduced, which, using data from users' profiles and their posts, created psychographic profiles for each of them. Researchers can use user posts in Facebook groups, comments and reactions to conduct emotional analysis and discover the intent of each user or group of users.
As was already mentioned on *promptcloud.com*[11], ‹‹ scraping is the automated collection of data (for example, using data collection software) from a website or from interfaces, providing features created for individuals. In general, the data scraping process: i) involves extracting data from various sources, including the World Wide Web ii) can be done at any scale iii) deduplication is not necessarily part of iv) requires crawl agent and parser to work. ››



A simple description of the data scraping process on Facebook
(image source: https://www.bestproxyreviews.com/facebook-scraper/**)**

Facebook provides APIs for collecting user profiles and content created by other users on the platform.

Essentially, the API is the acronym for Application Programming Interface, which is a software proxy that allows two applications to talk to each other. For example, every time someone uses an app (like Facebook), sends an instant message, or checks the weather on their phone, they use an API.

In the current research, an internet algorithm was used, called Facebook Scraper[12], which was implemented on public Facebook pages, without an API key, in order to collect data from the public Facebook profile of a user[13].

4.4.2 Application for user predictions

Through the application of the Facebook Scraper algorithm in the public profile of the user, the following elements were recorded in total: Posts = 38, Comments = 43689, Shares = 21088, Likes = 164087.
This user data was collected from the posts of the 10 most recent pages in his profile and saved in a csv file, the total number of posts the user has uploaded, the total number of comments, shares and likes the user has received in these posts by other Facebook users.

[10] https://www.bestproxyreviews.com/facebook-scraper/
[11] https://www.promptcloud.com/blog/data-scraping-vs-data-crawling/
[12] https://github.com/kevinzg/facebook-scraper#facebook-scraper
[13] The username is not listed due to Facebook's privacy policy.

	user_id	post_id	comments	shares	likes
1	12224403053	403974061056650	3013	1184	3701
2	12224403053	1249069962222351	2677	855	2848
3	12224403053	10159605780233054	89	0	2138
4	12224403053	10159604977038054	212	0	3743
5	12224403053	338687464578000	3300	1232	4387
6	12224403053	10159600987588054	351	0	3992
7	12224403053	333087361694659	5047	2103	7407
8	12224403053	10159596036333054	478	261	3446
9	12224403053	235959328164691	1627	928	3675
10	12224403053	10159590820763054	819	512	10824

4.4.2.1 User Personality Prediction

After collecting the data, the next step is to calculate the user's trust in the network. This is done initially, by applying the posts, comments, shares and likes, collected in the previous procedure, to the formulas presented in paragraphs 3.1.1 and 3.1.2, which calculate the User's Participation and Response Trust respectively.

The previous results are then used, together with the total number of posts uploaded by the user, as inputs to the Indirect Trust forecasting model presented in section 4.3.2. In this way, the level of trust between the user and the other friends on the network has now been successfully predicted.

The proposed model predicts that the Trust level is around 0.4999 and with r2_score being 0.999, it is stated that the regression model presented above fits almost perfectly.

The final step in predicting user personality is the application of user data: posts, comments, shares, likes, trust - as inputs to the AdaBoost classifier proposed in section 4.6. The goal is to predict whether the user rating for each of the Big Five Model personality traits is 1 or 0.

The results presented in **Table 8**, with Prediction Accuracy for each personality trait between 99% to 100% (**Table 6**), show that:

Table 8
The results of the user's personality

Personality Traits	SCORE	Deduction for Personality
Extraversion	1	extrovert, quite talkative, sociable and tends to enjoy human interactions with others in the community
Agreeableness	1	relatively polite, kind and sympathetic towards other people in the community
Conscientiousness	1	relatively careful, self-disciplined and tends to be dependable for others in the community
Neuroticism	0	calm, cheerful and has a strong control over his emotions
Openness to Experience	1	strong, tenacious and is constantly seeking new experience

4.4.2.2 User Interests Prediction

The final step of the research concerns the prediction of the user's interests, which is carried out in a similar way with the procedure of paragraph 4.8.2.1, i.e., initially with the application of the user data: posts, comments, shares, likes in the trust prediction model and then by applying them as inputs to the AdaBoost classifier proposed in section 4.6.

The goal this time is to predict the user's score from 1-5 for each of the interests presented in **Table 5**.

The results are presented in **Table 9**, with Prediction Accuracy for each of the interests / preferences as shown in **Table 7**, show that:

Table 9
The results of the user's interests

Interests	SCORE	Deduction for Interests
Music	5	The user enjoys listening to music a lot.
Movies	5	The user enjoys watching movies.
Musical Instruments	1	The user does not know or does not enjoy playing with musical instruments.
Fun with Friends	5	The user enjoys spending time and having fun with his friends.
Internet	5	The user enjoys spending a lot of time on the Internet.
Countryside/Outdoors	5	The user enjoys participating in outdoors activities at the countryside.
Sports	4	The user participates in sport activities.

5. Conclusions

The main goal of this work is to create accurate models from Facebook communities and groups, which will predict the dynamics in the relationships between users and will determine their social profile by identifying their personality traits and interests, in order to enable the classification of each Facebook user in the future, after first gathering the necessary data through a scraping algorithm.

The research presented a way to use the data collected by the Cheltenham dataset to measure the level of trust between users, first through the application of methods for calculating and forecasting the Indirect Trust globally and then directly between the pairs created between them.

The methods used in the work also prove that the Big Five personality traits of users, as well as their preferences, can be predicted from the public information available on Facebook.

Specifically, all 2538 users from the Facebook Cheltenham Groups dataset completed a personality test, while 1010 of them participated in another online interests survey and at the same time through the Facebook API, data were collected from their profiles. The questions included in these two tests, which made it possible to make an effective analysis for the parameters Extraversion, Agreeableness, Conscientiousness, Neuroticism, Openness to Experience, as well as for the main categories of interests / preferences - through application of certain methods, seem capable of determine a user's social behavior.

In order to draw safe conclusions with certainty, the proposals of the work had to be tested in different algorithms and methods and the results from each of them were compared according to their percentages accuracy.

After careful consideration, the regression model, which uses public data from the Cheltenham dataset and calculates the Indirect trust among users of a social network, gives the best results and predicts the level of trust of a user - whose public data was collected with the Facebook scraper - with an accuracy of 99%.

Also, the proposed classifier for calculating and predicting the score of personality and interests for each user of the dataset, the Ada Boosted Decision Tree algorithm, is performed with 99% - 100% accuracy for each of the 5 personality traits of Big Five model and with a variation between 60% - 90% for each interest category.

Judging by these results, it is easy to see that by applying a Facebook user's public data, first to the Indirect trust model, to measure the user's trust in a network and then to the proposed AdaBoost classifier, accurate results for the personality and preferences of the user are achieved.

References

[1] Shahid, Yousra Asim, Ahmad Kamran, Malik Basit Raza, Ahmad Raza. "A trust model for analysis of trust, influence and their relationship in social network communities." *Telematics and Informatics* (2018): 94-116. Document.

[2] Vedran Podobnik, Darko Striga, Ana Jandras, Ignac Lovrek. "How to Calculate Trust between Social Network Users?" *20th International Conference on Software, Telecommunications and Computer Networks (SoftCOM 2012)*. Ed. Vedran Podobnik. Split, Croatia: ResearchGate, 2012. Conference Paper. <https://www.researchgate.net/publication/256438314>.

[3] Greeshma Lingam, Rashmi Ranjan Rout, DVLN Somayajulu. "Learning automata-based trust model for user recommendations in online social networks." *Computers & Electrical Engineering* 66 (2017): 174-188. Document.

[4] Wei Chen, Simon Fong. "Social Network Collaborative Filtering Framework and Online Trust Factors: a Case Study on Facebook." *Research Gate* (2011). Publication. <https://www.researchgate.net/publication/220500799>.

[5] Frank E. Walter, Stefano Battiston, Frank Schweitzer. "Personalised and dynamic trust in social networks." *RecSys '09*. New York, USA: Association for Computing Machinery, New York, NY, United States, 2009. 197- 204. Research - Article.

[6] Golbeck, Jennifer. "Trust and nuanced profile similarity in online social networks." *ACM Transactions on the Web* 3.4 (2009): 1-33. Research - Article.

[7] Zhiyong Zhang, Kanliang Wang. "A trust model for multimedia social networks." *Social Network Analysis and Mining* 3 (2013): 969 - 979. Article.

[8] Alireza Souri, Shafigheh Hosseinpour, Amir Masoud Rahmani. "Personality classification based on profiles of social networks' users and the five-factor model of personality." *Human-centric Computing and Information Sciences* 8.24 (2018). Article.

[9] Michael M. Tadesse, Hongfei Lin, Bo Xu, Liang Yang. "Personality Predictions Based on User Behavior on the Facebook Social Media Platform." *IEEE Access* 6 (2018): 61959 - 61969. Article.

[10] Jennifer Golbeck, Cristina Robles, Karen Turner. "Predicting personality with social media." *CHI '11: CHI Conference on Human Factors in Computing Systems*. Vancouver BC Canada: Association for Computing MachineryNew YorkNYUnited States, 2011. 253 - 262. Document.

[11] Di Xue, Lifa Wu, Zheng Hong, Shize Guo, Liang Gao, Zhiyong Wu, Xiaofeng Zhong, Jianshan Sun. "Deep learning-based personality recognition from text posts of online social networks." *Applied Intelligence* 48 (2018): 4232 - 4246. Article.

[12] Buettner, Ricardo. "Predicting user behavior in electronic markets based on personality-mining in large online social networks." *Electronic Markets* 27 (2017): 247 - 265. Research Paper.

[13] https://www.simplypsychology.org/big-five-personality.html

[14] https://python.plainenglish.io/adaboost-classifier-in-python-8d34a9f20459

[15] https://www.criteriacorp.com/resources/glossary/big-five-personality-traits

[16] https://towardsdatascience.com/machine-learning-part-17-boosting-algorithms-adaboost-in-python-d00faac6c464

[17] https://en.wikipedia.org/wiki/Linear_regression#Generalized_linear_models

[18] http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm

[19] http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm

[20] https://www.w3schools.com/

[21] https://www.promptcloud.com/blog/data-scraping-vs-data-crawling/

[22] https://www.bestproxyreviews.com/facebook-scraper/

[23] http://www.sthda.com/english/wiki/ggplot2-scatter-plots-quick-start-guide-r-software-and-data-visualization