POLITECNICO DI TORINO

Master's Degree in Communication and Computer Networks Engineering

Master's Degree Thesis

Methodologies to derive subjective risk for web tracking

Supervisors

Prof. Marco MELLIA

Dott. Martino TREVISAN

Dott. Luca VASSIO

Candidate

Vito Daniele GAMBINA

April 2022

Abstract

This work aims to increase the users' awareness while browsing the internet, introducing them to today's tracking ecosystem and derive their perceived risk to assign to websites a subjective risk indicator score. In the digital era where we live, in almost all families it is possible to find a device, such as smartphone, pc, tablet, capable of connecting to the internet and allowing them to visit dozens of websites every day. During their daily online activity, people are unaware to encounter dozens and dozens of web trackers, that, nowadays, represent the most widespread threat to our privacy, allowing the slow and constant accumulation of different kinds of online data in order to build users profiles and to customize targeted ads or other things. For this reason focusing on privacy online and data security is increasingly important and provide users, during their navigation, an indicator of websites risk may be a first step to improve their online experience. In this thesis, a simply survey was developed where some tracking features of the most important websites were presented to some normal web users in order to derive their perceived risk. The results gathered from the survey conducted were analyzed through a machine learning algorithm useful for this thesis purpose. The choice fell on the linear regression algorithm, one of the most basic tools in the area of machine learning for prediction. This was used to estimate the relationships between the objective tracking data and the final risk score indicated by users for each website with the final purpose to construct a model able to predict this score also for other websites. The linear regression model built performs very well reaching very good level of accuracy and shows that machine learning algorithms can be considered for this kind of situation. The results obtained through this thesis work provide users with a better awareness in controlling their data and provide a new point of view for future studies on the web tracking ecosystem.

Contents

Chapter 1

Introduction

1.1 Web tracking and PimCity project

Web tracking is the most widespread threat to users privacy, with this practice web trackers are able to collect users' data and their web browsing behavior.

In recent years, the internet is playing an increasingly important role within our society, simplifying and speeding up the search for information and the performance of many tasks, such as reading a newspaper, shopping or even working. In particular, with Covid-19 pandemic, where movements are limited, many people have started using the internet more frequently, discovering the simplicity to do things from home and with a few clicks. In parallel to this digital trend, companies need to meet customer demands in short terms and many of them have identified web tracking technologies as a solution. These companies can be called web trackers and, collect, everyday, information about users' needs and habits in order to build users profiles and to more efficiently target products and services to their customers or to sell them to ads companies, acting as data brokers. If on the one hand, web tracking can lead a market advantage for many companies, on the other can undermine the security of the user's personal data, and for this reason it has become a current and delicate topic to be explored.

When a user browses on a website, he may be tracked by two kind of web trackers: first party domain, which is the website the user is visiting, or by a third-party domain, different from the visited websites. The reasons are many, in fact, in addition to advertising purposes, web tracking may be used either by law enforcement agencies to spy on individuals and solve crimes, either for web analytics purposes to evaluate the performance of a website, or for usability tests, testing how a website's design is easy to use. In most of these cases, users are unaware of this tracking ecosystem which very often acts in the shadows by relying on host of solutions to identify a user.

The most used and known technique is using cookies, pieces of code embedded onto users' devices or browser to recognize them on subsequent visits, allowing to save their preferences or past purchases in order to customise the user's browsing experience (first party trackers) or to track their online activity on other websites to deliver targeted ads (third party trackers). For what concerns third party domains, in addition, it is possible to encounter more subtle tracking techniques such as, beacons or fingerprinting.

Beacons, also known as Web beacons, are single-pixel (graphics interchange format, GIF) image tags in hypertext markup language (HTML) documents placed on a Web site or in an e-mail message to monitor user behavior. While users have the option to accept or decline cookies, beacons are placed within HTML as a small, imperceptible graphic file, often transparent because it is the same color as the background. When a user opens the page or email where such an image is embedded, they might not see the image, but their web browser or email reader automatically downloads the image, requiring the user's computer to send a request to the host company's server, where the source image is stored. This request provides identifying information about the computer (e.g. IP address, time of request, type of web browser), allowing the host to keep track of the user.

Web fingerprinting is an advanced technique for gathering information about users when they browse the Internet. It is a persistent tracking technique which does not require a tracker to set any state in the user's browser, but it attempts to identify users by a combination of the device's properties. Its deployment is aimed to uniquely identify users without relying on cookies or other kinds of client-side state. Based on information obtained from the browser and device, fingerprinting practice builds precise signatures to uniquely re-identify them across different web services. Fingerprints can be obtained on-the-fly by injecting specialized JavaScript code, which the browser executes in a transparent way to the user. Nowadays, fingerprinting is a common practice and it has been widely studied by the research community because of its deep consequences, in fact, it allows trackers to be more precise at recording users' online behavior, putting users' privacy seriously at risk.

The continuous spread of new and increasingly pervasive tracking techniques has, therefore, raised many concerns about privacy, leading Europe to introduce new legislation. In 2018 became enforceable the GDPR (General Data Protection Regulation) [1], a regulation in EU law on data protection and privacy in the European Union (EU) and the European Economic Area (EEA), with primary aim to enhance individuals' control and rights over their personal data and to simplify the regulatory environment for international business.

In this moment of the history where the web has, therefore, become a large data market and people are not aware to blindly provide their data everyday for free access to services, the EU founded the PimCity project [2], that offers tools to change this scenario. PimCity borns, therefore, with the aim to increase transparency and provide users with control over their data. The idea of the project is to implement a PIMS (Personal Information Management Systems) that can be thought as a software layer between end users and services, responsible for releasing data from the former towards the latter, in a controlled manner. In addition, PimCity designs and deploys novel mechanisms to increase users' awareness, such as Personal Data Avatar (PDA), the Transparency Tags (TT). The PDA is the interface between the user and the services, where the user becomes the only owner of her data and acquires the freedom and power of deciding which data to share with which service. Transparency Tags (TT) is the analogue to a Nutrition Label for food which provides the information about the ingredients, their provenience, intolerance risks, etc. of food, in particular, it communicates in an easy to understand way the nature of the web service the user is accessing to. The idea is to have a label for every website that summarizes into scores the potential privacy risk associated to each of them starting from some information such as its owner, its purpose, the personal data it collects, etc.

This project can be the first step to a safer web browsing experience, shedding light on the shadow of online tracking and increasing users' awareness of their data and privacy.

1.2 Goal

This work was born with the idea of making a contribution to the construction of Transparency Tags (TT). To provide users with a more complete view of the nature of the website they are visiting, in addition to the objective parameters about the number of trackers present, the techniques used, etc., this thesis work wants to add a subjective component about the users perception of risk. For this purpose, after a careful study of the objective characteristics of some of the most visited websites in Italy, Great Britain, Spain and in the world, thanks to the dataset provided by Ermes Company [3], it was possible to carry out an articulated survey in order to grasp the different reactions of the users once discovered the world of web tracking behind the sites analyzed. With the information obtained from the survey it was possible to notice some different reactions among the interviewed population, confirming how certain circumstances can affect the perception of some aspects of the Web world and providing us possible ideas for future work. Furthermore, by combining the objective data (previously mentioned) and the new subjective data through an appropriate machine learning technique (linear regression) it was possible to obtain a subjective score regarding the potential risk of websites, that can be used to construct the Transparency Tags (TT).

Chapter 2

State of the art

The research community's interest in web tracking comes relatively recently in the history of web, indeed, despite the earliest tracker observed dates back to 1996, the earliest measurement studies began in 2005, with most coming after 2009. According to studies conducted by the university of Washington, websites contact an increasing number of third parties over time: in early 2000s, only the 5% of the 500 most popular sites contacted at least 5 separate third parties, whereas nearly 40% in 2016 [4]. This significant increase explains why studies on the world of web tracking have intensified only in the last years.

Hand in hand with the spread of the internet in the world, we find therefore the evolution of the web tracking ecosystem that causes concern for the privacy of users. For this reason, Krishnamurthy and Wills, between October 2005 and September 2008, were among the first to examine well-known web tracking techniques and their degree of penetration in popular Web sites. They have examined the role of cookies and JavaScript and the potential of users aggregate data for companies. From their work is emerged that, by purchasing behavioral data from the past, the acquiring companies are able to get a broader idea about the behavior of users over time which can be helpful to predict future trends and to convert user-neutral data into identifiable user profile [5].

For this reason, online tracking has proved to be a fundamental tool for companies to meet customer demands and it has therefore often been described as an "arms race" that also includes other more advanced and more pervasive techniques than cookies. In a paper of Princeton University [6], in fact, tracking techniques that are hard to control, hard to detect and resilient to blocking or removing as Canvas fingerprinting, Evercookies and Cookie syncing have been studied and described. The former is a mechanism that uses the browser's Canvas API to draw invisible images and extract a persistent, long-term fingerprint without the user's knowledge. There doesn't appear to be a way to automatically block canvas fingerprinting without false positives that block legitimate functionality. Evercookies actively circumvent users' deliberate attempts to start with a fresh profile by abusing different browser storage mechanisms to restore removed cookies. Cookie syncing allows different trackers to share user identifiers with each other.

To better understand the web tracking ecosystem, also in 2016, Englehardt and Narayanan [7] performed the largest and most detailed measurement of online tracking conducted to date, based on a crawl of the top 1 million websites, with over 90 million requests. They found out that the total number of third parties present on at least two first parties is over 81,000, but only 123 of these 81,000 are present on more than 1% of sites. The level of tracking also varies on different categories of sites, news sites being the ones with the higher numbers of third parties, while sites belonging to government organizations, universities, and non-profit entities tend to have lower numbers.

The European Union (EU) against these online tracking technologies introduced a first set of regulations in 2002, the ePrivacy Directive, that aimed, among other things, to make online tracking mechanisms explicit to increase privacy awareness among users. This directive became mandatory only in 2013 and it mandates websites to ask for informed consent before using any kind of profiling technology, e.g., cookies and most of European websites embed a “Cookie Bar”, the most visible effect of the regulation. However, a study conducted in [8] testified that a wide fraction of websites does not respect the the Cookie Law set up by the ePrivacy Directive, with a few popular third parties causing such violations.

In this context full of data privacy pitfalls, users' concerns about online tracking are difficult to capture because they change under specific circumstances. In fact, this is confirmed by studies of Carnegie Mellon University in 2016 [9], from where it is emerged that users are less comfortable with the invisible outcomes of tracking (price discrimination, revenue for web sites, etc.) than with more noticeable outcomes (ads, customization, etc.), and that users commonly base their tracking preferences on specific properties of first-party websites, such as the topic of the site and frequency of visits.

In 2018, Jaspreet Bhatia and Travis D. Breaux, always belonging Carnegie Mellon University, introduced an empirical framework in [10] that consists of factorial vignette surveys that can be used to measure the effect of different factors and their levels on privacy risk. They reported a series of experiments to measure perceived privacy risk using a proposed framework, which are based on expressed preferences, and which defined as an individual's willingness to share their personal data with others given the likelihood of a potential privacy harm. These experiment founded that participants are more willing to share their information when they perceive the benefits of sharing.

On this line, other studies are still being carried out in order to provide a complete and updated view of the actual use of new advanced techniques on the web, as done in [11] with the final goal to find the right countermeasures to be able to guarantee more transparency in the web world and consequently ensure greater security of users data during their web navigation. This thesis will try to make a contribution to achieve this last goal, proposing a different approach than the studies conducted up to now, where people will be the protagonists and will express a subjective opinion once in contact with the world of web tracking.

Chapter 3

Study and Analysis of the Ermes Dataset

An important step of this thesis work is the characterization of data made available by Ermes company. In this phase I studied and analyzed these data in order to provide web tracking information to the users of the survey with a simpler and more understandable key. This process was done through Python which allowed me, thanks to the development of a simple code, to do this job more efficiently by quickly gathering the most important information from text files with a large amount of data.

The dataset provided is composed by a list of files that include statistics that the Ermes company computed thanks to a crawling campaign carried out on the first 10k websites of the Tranco ranking [12], visiting for each domain, included in this list, its homepage (if reachable) and up to 20 sub-pages (if any). From this campaign web tracking information was extracted using various OSINT (Open Source INTelligence) sources and some proprietary algorithms and these data were then processed with Spark tool to obtain the results of the files that make up our dataset. From this dataset, it is possible to identify two kind of files:

- **per-websites stats files**, that includes all files containing information on web tracking operations found in all the websites analyzed.

- **per-trackers stats files**, on the other hand, includes all the files containing information, mostly statistics, on the trackers, in order to understand their danger to the privacy of users.

From the first category of files it is possible to obtain immediately recognizable statistical data that can give a first idea of the risk that it is possible to encounter on the site in question. In particular, the following three files were considered:

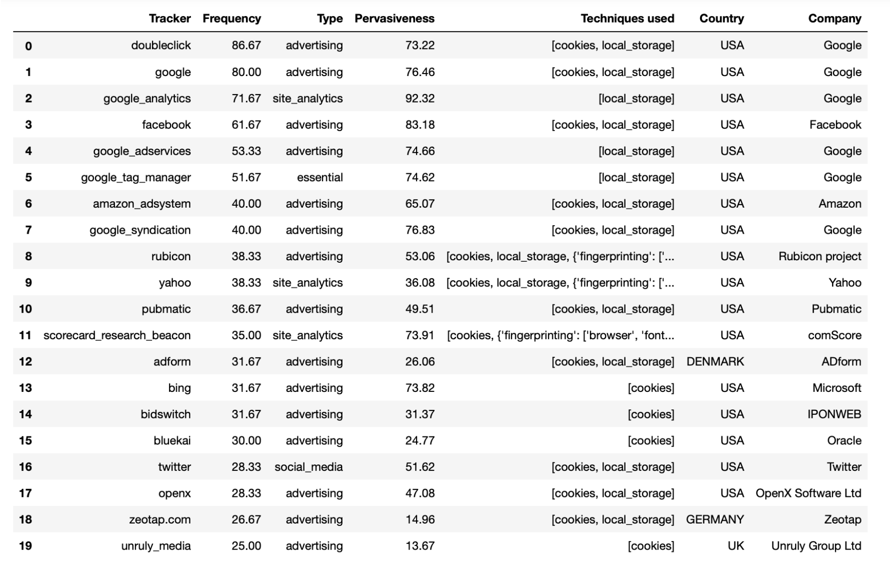
- **trackers-by-visit-url.txt**, that includes, for each website, the list of trackers that have been found on it, with some information such as its domain, the type and its owner company.

- **avg-content-length.txt**, that includes, for each website, the average of the sum of tracker-related content lengths on the website's homepage and subpages.

- **avg-tracking-requests.txt**, that includes, for each website, the average number of tracker-related HTTP requests on the website's homepage and subpages.

From these files, above mentioned, it is possible to show to users, for each websites, a first layer of information about the web tracking ecosystem around the website analyzed such as the **number of trackers**, **the content length**, in terms of byte, related to trackers present in the website and **the average number of contact** between website and its trackers when a user start to visit the website.

To go more deeper into the world of web tracking behind these websites and to understand the type of trackers present into them, the second category of files was used. To make up for some missing data in the files, some individual searches were made, in order to arrive at a complete picture of the trackers. However, our attention fell only on the top 20 most frequent trackers in our database, choosing to not overload the user with a myriad of information on all trackers but to focus his attention only on a smaller circle of trackers. These data can be summarized by a simple dataset constructed with Python:



	Tracker	Frequency	Type	Pervasiveness	Techniques used	Country	Company
0	doubleclick	86.67	advertising	73.22	[cookies, local_storage]	USA	Google
1	google	80.00	advertising	76.46	[cookies, local_storage]	USA	Google
2	google_analytics	71.67	site_analytics	92.32	[local_storage]	USA	Google
3	facebook	61.67	advertising	83.18	[cookies, local_storage]	USA	Facebook
4	google_adservices	53.33	advertising	74.66	[local_storage]	USA	Google
5	google_tag_manager	51.67	essential	74.62	[local_storage]	USA	Google
6	amazon_adsystem	40.00	advertising	65.07	[cookies, local_storage]	USA	Amazon
7	google_syndication	40.00	advertising	76.83	[cookies, local_storage]	USA	Google
8	rubicon	38.33	advertising	53.06	[cookies, local_storage, {'fingerprinting': ['...	USA	Rubicon project
9	yahoo	38.33	site_analytics	36.08	[cookies, local_storage, {'fingerprinting': ['...	USA	Yahoo
10	pubmatic	36.67	advertising	49.51	[cookies, local_storage]	USA	Pubmatic
11	scorecard_research_beacon	35.00	site_analytics	73.91	[cookies, {'fingerprinting': ['browser', 'font...	USA	comScore
12	adform	31.67	advertising	26.06	[cookies, local_storage]	DENMARK	ADform
13	bing	31.67	advertising	73.82	[cookies]	USA	Microsoft
14	bidswitch	31.67	advertising	31.37	[cookies]	USA	IPONWEB
15	bluekai	30.00	advertising	24.77	[cookies]	USA	Oracle
16	twitter	28.33	social_media	51.62	[cookies, local_storage]	USA	Twitter
17	openx	28.33	advertising	47.08	[cookies, local_storage]	USA	OpenX Software Ltd
18	zeotap.com	26.67	advertising	14.96	[cookies, local_storage]	GERMANY	Zeotap
19	unruly_media	25.00	advertising	13.67	[cookies]	UK	Unruly Group Ltd

Figure 3.1: Dataset of the top 20 trackers' features

In the table shown above are summarized all the characteristics of 20 trackers more frequent in the database analyzed (as shown in the first column "Frequency"). In fact, for every 20 trackers, it is possible to identify:

- **Frequency**, the share (in %) of visited websites where web tracker has been encountered on.
- **Type**, represent the type of service trackers provide to the owner and trackers can be categorized in the following way:
    1. *advertising*, provides advertising or advertising-related services such as data collection, behavioral analysis or re-targeting;
    2. *comments*, enables comments sections for articles and product reviews;
    3. *customer interaction*, includes chat, email messaging, customer support, and other interaction tools;
    4. *essential*, includes tag managers, privacy notices, and technologies that are critical to the functionality of a website;
    5. *pornvertising*, delivers advertisements that generally appear on sites with adult content;
    6. *site analytic*, collects and analyzes data related to site usage and performance;
    7. *social media*, integrates features related to social media sites;
    8. *audio video player*, enables websites to publish, distribute, and optimize video and audio content;
    9. *CDN (Content Delivery Network)*, content delivery network that delivers resources for different site utilities and usually for many different customers;
    10. *misc (Miscellaneous)*, this tracker does not fit in other categories;
    11. *hosting*, this is a service used by the content provider or site owner.
    12. *unknown*, this tracker has either not been labelled yet, or we do not have enough information to label it.
- **Pervasiveness**, the percentage of website's subpages where web tracker has been found on.
- **Techniques used**, list of all web tracking techniques used by the web tracker:
    - *HTTP cookies*, are code and information embedded onto a user's device by a website when the user visits the website.

  - *Local storage*, very similar to the characteristics declared for cookies, but in this case the code and information embedded onto a user's device does not have an expiration date and therefore must be deleted by the user.
  - *device fingerprint or machine fingerprint or browser fingerprint*, is a technique that allows collecting basic information on the web browser's configuration to identify, in whole or in part, individual users or devices even when cookies are disabled. The assimilation of this information into a single string allows creating a fingerprint of the device.

- **Country**, the country where is based the web tracker.
- **Company**, the company which owns the web tracker.

Thanks to the detailed analysis of the available Dataset and the further research done and described in this paragraph, it was possible to start developing the ideas for the survey to be able to show users the obtained data in a simple way.

Chapter 4

Survey

At this point I started to develop a survey in order to introduce people to the web tracking world and understand and capture their different reactions and perceptions by looking at the data collected on the most popular sites in the world.

4.1 Survey Development

The survey was developed through LimeSurvey, a free and open source on-line statistical survey web app that enables users using a web interface to develop and publish on-line surveys, collect responses, create statistics, and export the resulting data to other applications. With this tool, in fact, it was possible to develop and customize our survey thanks to the possibility to use rich text in questions and messages, integrate images and videos and modify the layout and design of the survey with an HTML editor.

The main idea is to collect aware feedback from the users interviewed and for this reason it was important to introduce them into the web tracking world with small steps. The questionnaire, in fact, is composed by:

- a *preliminary section*, composed by three pages, where the interviewee is asked to answer to some personal and technical questions to understand people's background and prepare them to the next section.

- a *risk evaluation section*, the real important part of the survey, where the interviewee is asked to value 5 different websites on the basis of his reactions and subjective risk perceived after looking to their web tracking statistics.

Furthermore, the survey was developed both in Italian, English and Spanish to make it closer to a great variety of users and to get more varied answers. The version can be chosen at the beginning of the survey on its first welcome page. For this reason three different sets of 25 websites were built selecting 25 Italian websites, 25 Britain websites and 25 Spanish websites, from which 5 websites will be randomly taken for each interviewee. Each set is composed by 5 different categories, each including 5 specific websites, with different characteristics and varied web tracking data. The categories chosen, in fact, try to cover most of the possible areas of user interest:

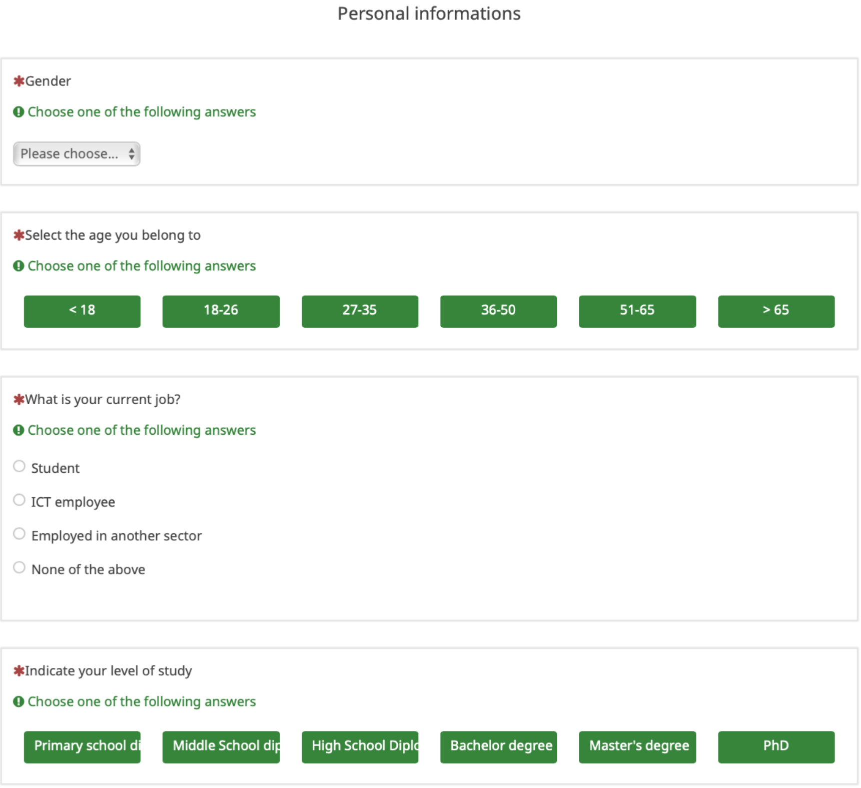
1. newspapers category;
2. consumer electronics category;
3. TV movies and streaming category;
4. e-commerce and shopping category;
5. miscellaneous category of music, online banks and programming.

4.1.1 Preliminary section

Going into detail, the first section is a sort of introduction to the more substantial section of the risk assessment of websites, but it is nevertheless a section that should not be underestimated because it is important in order to characterize the survey participants and to reach , at the end, a more detailed conclusion.

The preliminary section, as previously mentioned, is composed by three pages that are organized in the following way:

1. a first one where the interviewee is asked to answer to four personal questions about gender, age, current job and level of education fig. 4.1, in order to have an idea of the type of people who participated in the survey and keep track of the differences in their answers;



Personal informations

*Gender

Choose one of the following answers

Please choose...

*Select the age you belong to

Choose one of the following answers

< 18 | 18-26 | 27-35 | 36-50 | 51-65 | > 65

*What is your current job?

Choose one of the following answers

○ Student

○ ICT employee

○ Employed in another sector

○ None of the above

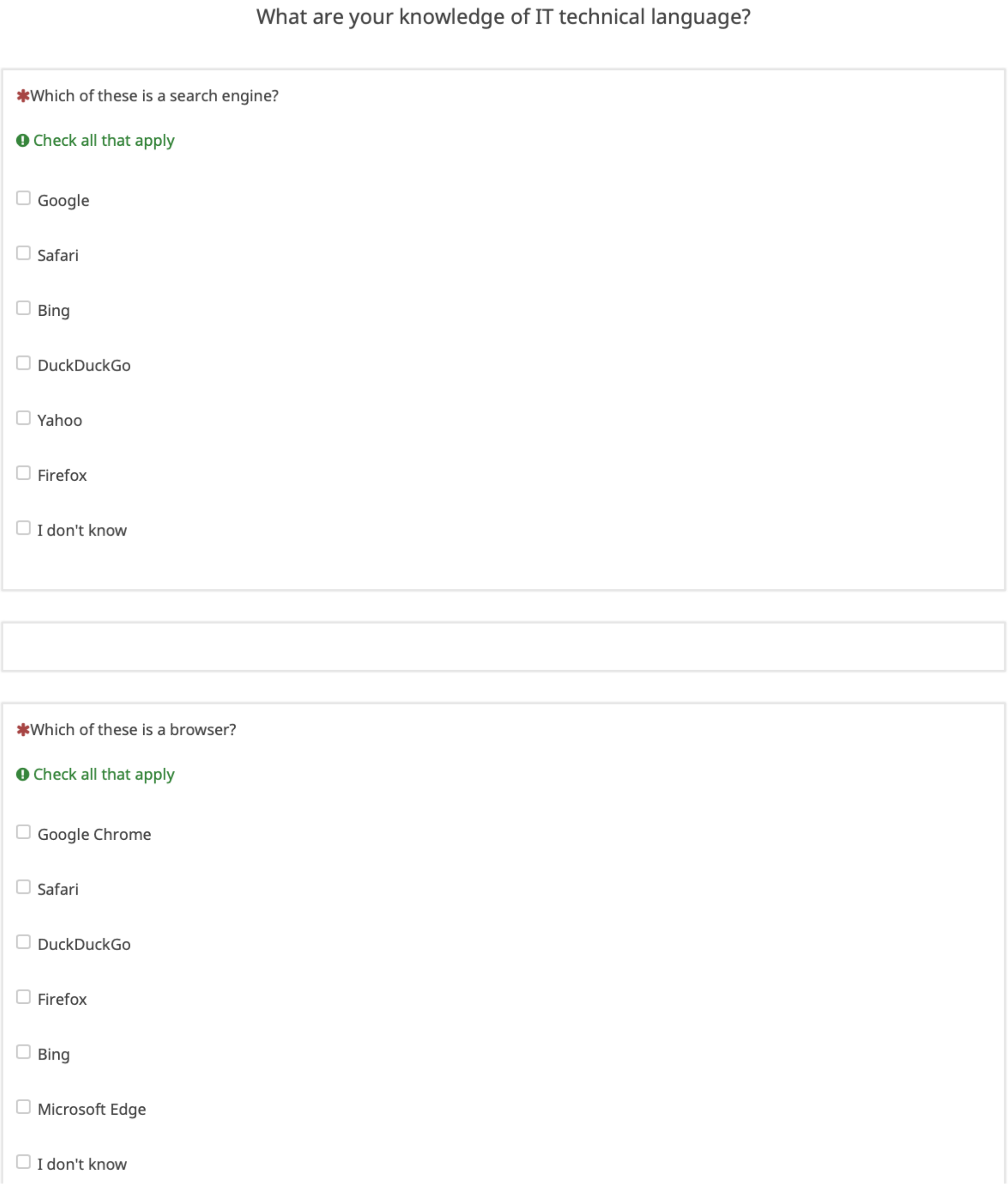
*Indicate your level of study

Choose one of the following answers

Primary school di | Middle School dip | High School Diplo | Bachelor degree | Master's degree | PhD

Figure 4.1: Personal Informations

2. the second one where he is asked to answer to five question on basic terms of computer language and web tracking fig. 4.2 and fig 4.3, in order to understand his level of knowledge in this field.



What are your knowledge of IT technical language?

*Which of these is a search engine?

Check all that apply

□ Google

□ Safari

□ Bing

□ DuckDuckGo

□ Yahoo

□ Firefox

□ I don't know

*Which of these is a browser?

Check all that apply

□ Google Chrome

□ Safari

□ DuckDuckGo

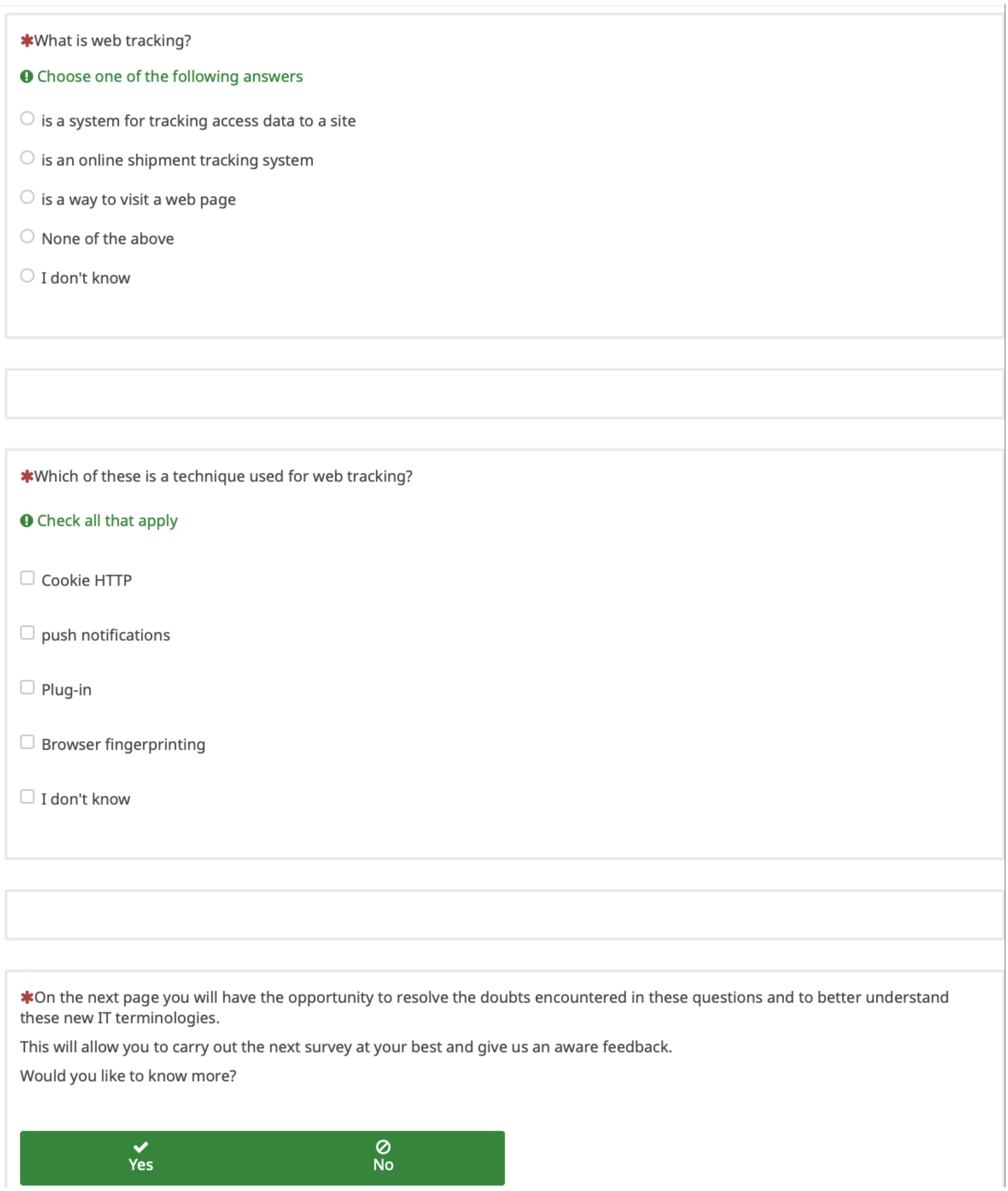
□ Firefox

□ Bing

□ Microsoft Edge

□ I don't know

Figure 4.2: First part of theory questions



✱What is web tracking?

❶ Choose one of the following answers

○ is a system for tracking access data to a site

○ is an online shipment tracking system

○ is a way to visit a web page

○ None of the above

○ I don't know

✱Which of these is a technique used for web tracking?

❶ Check all that apply

□ Cookie HTTP

□ push notifications

□ Plug-in

□ Browser fingerprinting

□ I don't know

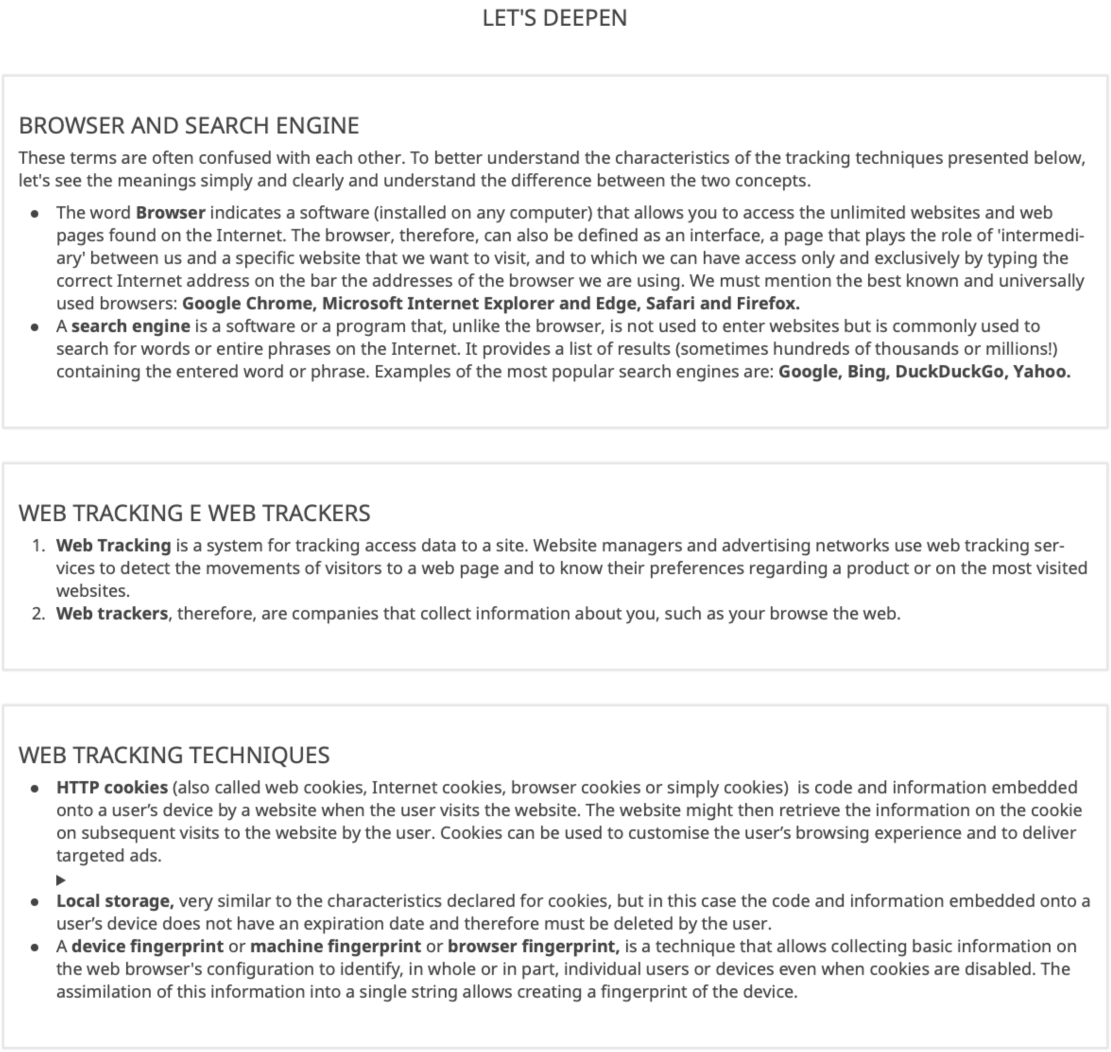
✱On the next page you will have the opportunity to resolve the doubts encountered in these questions and to better understand these new IT terminologies.

This will allow you to carry out the next survey at your best and give us an aware feedback.

Would you like to know more?

Figure 4.3: Second part of theory questions

3. optional third section to which the interviewee can decide to access or not and where he can read and learn simple concepts about web tracking fig. 4.4, discovering the correct answers of the previous section and better prepare himself for the next section.



LET'S DEEPEN

BROWSER AND SEARCH ENGINE

These terms are often confused with each other. To better understand the characteristics of the tracking techniques presented below, let's see the meanings simply and clearly and understand the difference between the two concepts.

- The word **Browser** indicates a software (installed on any computer) that allows you to access the unlimited websites and web pages found on the Internet. The browser, therefore, can also be defined as an interface, a page that plays the role of 'intermediary' between us and a specific website that we want to visit, and to which we can have access only and exclusively by typing the correct Internet address on the bar the addresses of the browser we are using. We must mention the best known and universally used browsers: **Google Chrome, Microsoft Internet Explorer and Edge, Safari and Firefox.**
- A **search engine** is a software or a program that, unlike the browser, is not used to enter websites but is commonly used to search for words or entire phrases on the Internet. It provides a list of results (sometimes hundreds of thousands or millions!) containing the entered word or phrase. Examples of the most popular search engines are: **Google, Bing, DuckDuckGo, Yahoo.**

WEB TRACKING E WEB TRACKERS

1. **Web Tracking** is a system for tracking access data to a site. Website managers and advertising networks use web tracking services to detect the movements of visitors to a web page and to know their preferences regarding a product or on the most visited websites.
2. **Web trackers**, therefore, are companies that collect information about you, such as your browse the web.

WEB TRACKING TECHNIQUES

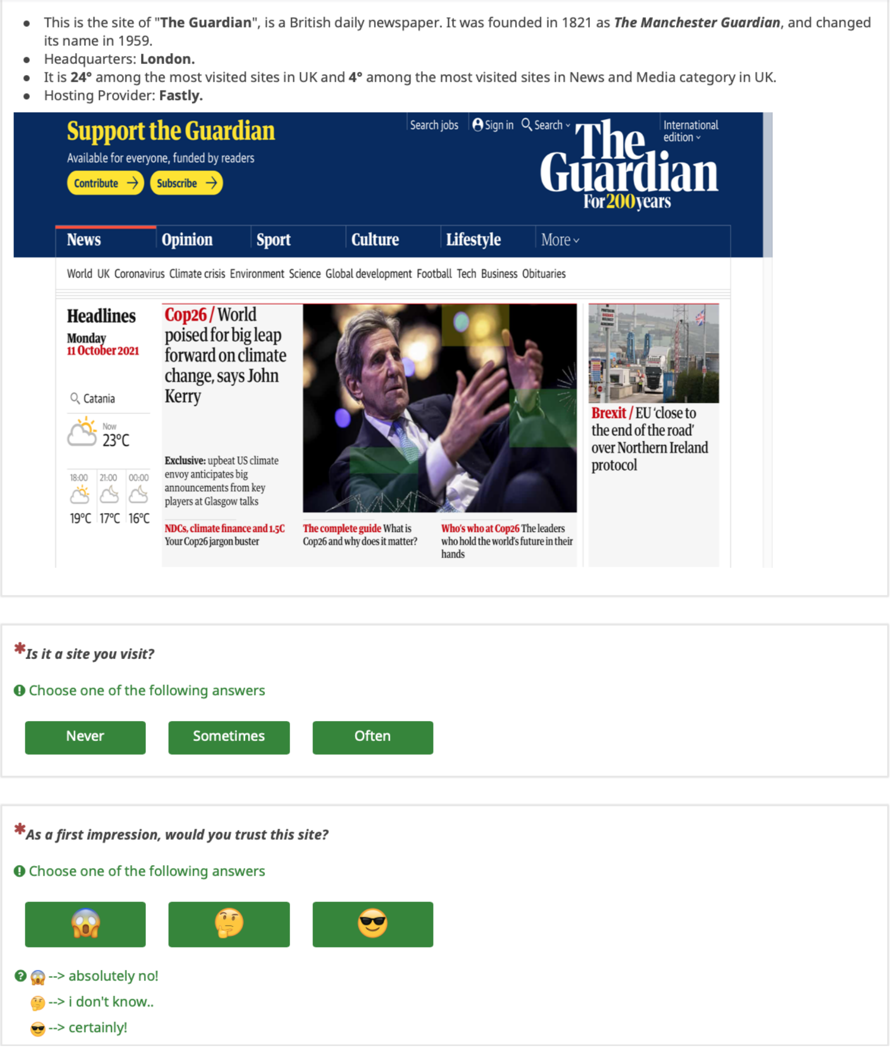
- **HTTP cookies** (also called web cookies, Internet cookies, browser cookies or simply cookies) is code and information embedded onto a user’s device by a website when the user visits the website. The website might then retrieve the information on the cookie on subsequent visits to the website by the user. Cookies can be used to customise the user’s browsing experience and to deliver targeted ads.
  ▶
- **Local storage,** very similar to the characteristics declared for cookies, but in this case the code and information embedded onto a user’s device does not have an expiration date and therefore must be deleted by the user.
- A **device fingerprint** or **machine fingerprint** or **browser fingerprint,** is a technique that allows collecting basic information on the web browser's configuration to identify, in whole or in part, individual users or devices even when cookies are disabled. The assimilation of this information into a single string allows creating a fingerprint of the device.

Figure 4.4: In-depth Section

4.1.2 Risk evaluation section

The second and last section of the survey includes the evaluation of the risk of 5 websites, chosen randomly from the set of 25 selected websites. Each interviewee is therefore asked to express their impression on these websites in three different moments:

1. in the first page where the interviewee gives a **first impression** on the websites looking only at their homepage and a brief description, as it possible to see in the example of "The Guardian", a british newspaper, extrapolated from the first category of the british dataset in fig. 4.5.



- This is the site of "**The Guardian**", is a British daily newspaper. It was founded in 1821 as ***The Manchester Guardian***, and changed its name in 1959.
- Headquarters: **London.**
- It is **24°** among the most visited sites in UK and **4°** among the most visited sites in News and Media category in UK.
- Hosting Provider: **Fastly.**

Support the Guardian
Available for everyone, funded by readers
Contribute →
Subscribe →
Search jobs
Sign in
Search
The Guardian
For 200 years
International edition
News
Opinion
Sport
Culture
Lifestyle
More
World UK Coronavirus Climate crisis Environment Science Global development Football Tech Business Obituaries
Headlines
Monday 11 October 2021
Catania
Now 23°C
18:00 21:00 00:00
19°C 17°C 16°C
Cop26 / World poised for big leap forward on climate change, says John Kerry
Exclusive: upbeat US climate envoy anticipates big announcements from key players at Glasgow talks
NDCs, climate finance and 1.5C Your Cop26 jargon buster
The complete guide What is Cop26 and why does it matter?
Who's who at Cop26 The leaders who hold the world's future in their hands
Brexit / EU 'close to the end of the road' over Northern Ireland protocol

****Is it a site you visit?***

Choose one of the following answers

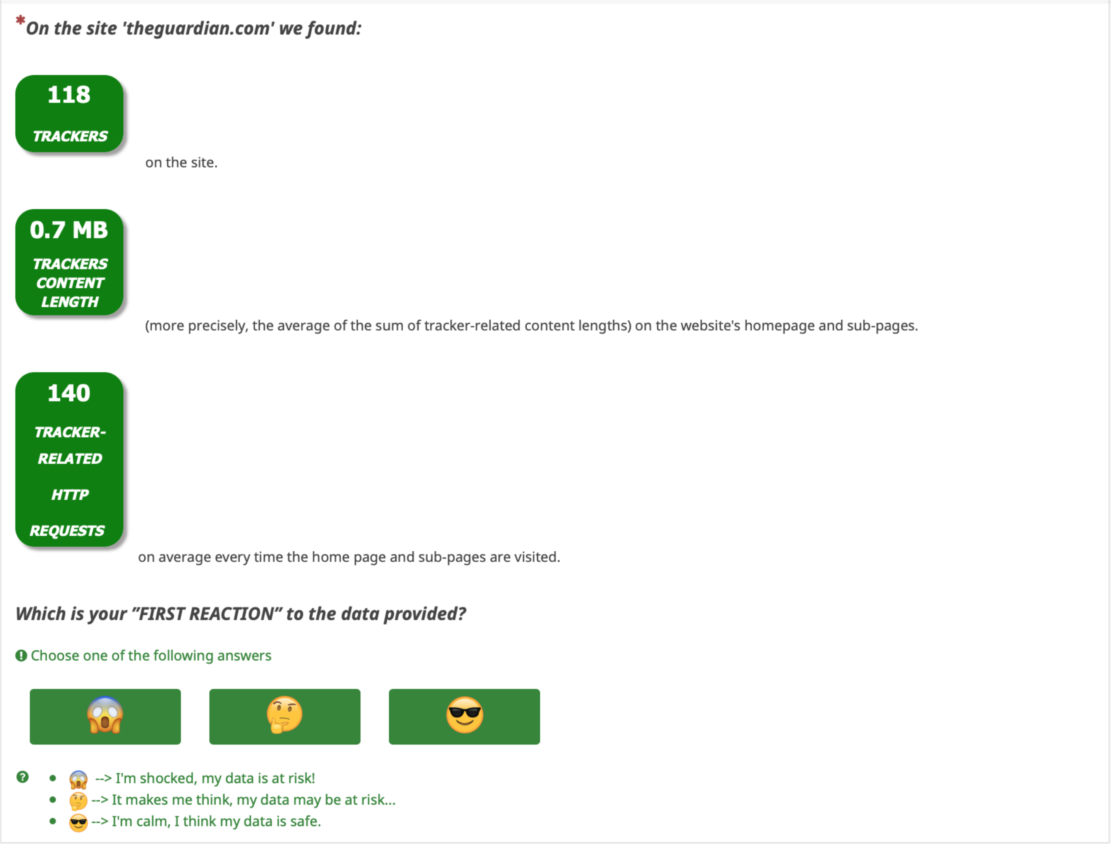
****As a first impression, would you trust this site?***

Choose one of the following answers

😱 --> absolutely no!
🤔 --> i don't know..
😎 --> certainly!

Figure 4.5: First page of the Guardian survey

2. in the second page where the interviewee gives a **first reaction** after looking at the web tracking data of the website, extrapolated, as mentioned in the paragraph 3, from per-websites stats files, as it possible to see in fig. 4.6.



****On the site 'theguardian.com' we found:***

118
TRACKERS

on the site.

0.7 MB
TRACKERS CONTENT LENGTH

(more precisely, the average of the sum of tracker-related content lengths) on the website's homepage and sub-pages.

140
TRACKER-RELATED HTTP REQUESTS

on average every time the home page and sub-pages are visited.

Which is your "FIRST REACTION" to the data provided?

Choose one of the following answers

😱 🤔 😎

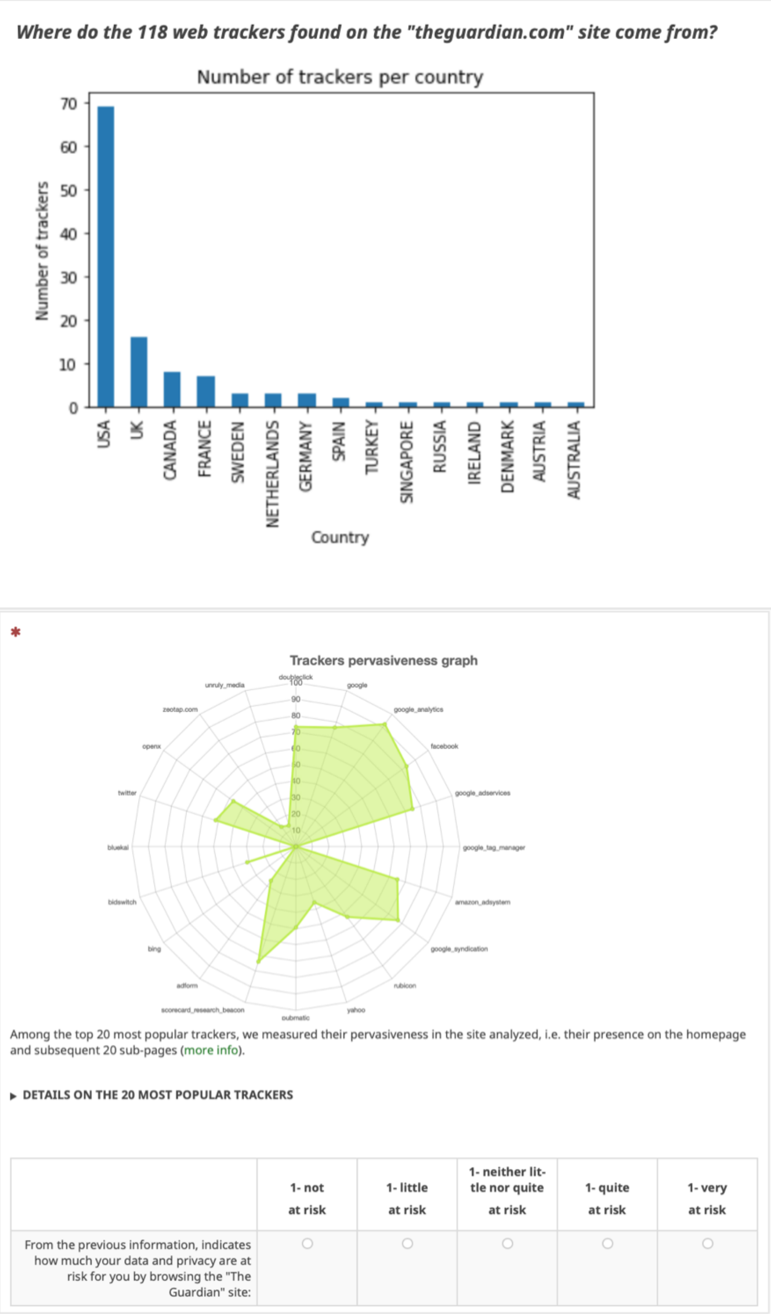
- 😱 --> I'm shocked, my data is at risk!
- 🤔 --> It makes me think, my data may be at risk...
- 😎 --> I'm calm, I think my data is safe.

Figure 4.6: Second page of the Guardian survey

3. in the third page where the interviewee give a final **overall valuation**, relating to his perceived subjective risk, in a scale from 1 to 5, where 1 corresponds to a null perceived risk and 5, instead, a very high perceived risk. On this page the numbers, relating to the trackers, provided on the second page are analyzed and detailed, with the aim to provide to the interviewees a complete view of the websites analyzed.

   For this purpose they have been provided to the user:

   - a bar plot, where it is possible to know the country of origin of the trackers encountered on the website;
   - a radar chart, where it is possible to understand the degree of pervasiveness of the top 20 trackers present in the website. With this kind of graph it is possible to visually understand the impact of the trackers on the website analyzed. Simply, the interviewee knows and immediately recognizes that the larger the area, the greater the possibility of meeting these trackers on the homepage and sub-pages of the analyzed site, in this example 'theguardian.com' (fig.4.7);
   - details on the top 20 trackers as their purpose, their type, techniques used and frequency(fig. 4.8), extrapolated in this case from per-trackers stats files.



Where do the 118 web trackers found on the "theguardian.com" site come from?

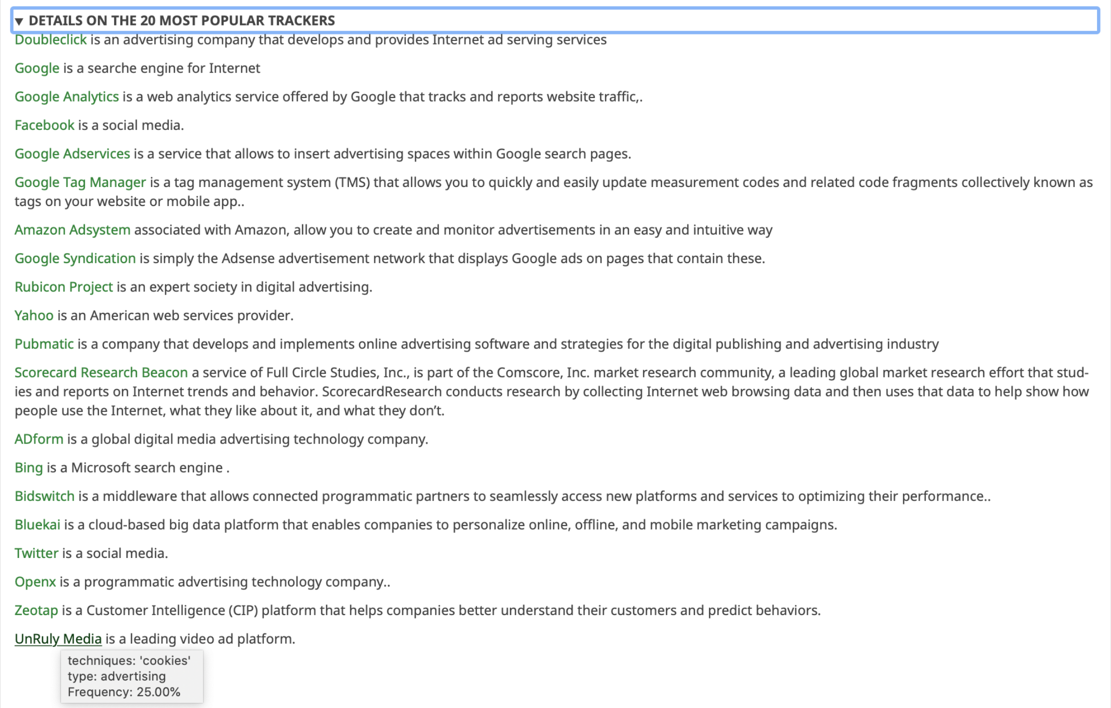
*

Among the top 20 most popular trackers, we measured their pervasiveness in the site analyzed, i.e. their presence on the homepage and subsequent 20 sub-pages (more info).

▸ DETAILS ON THE 20 MOST POPULAR TRACKERS

	1- not at risk	1- little at risk	1- neither little nor quite at risk	1- quite at risk	1- very at risk
From the previous information, indicates how much your data and privacy are at risk for you by browsing the "The Guardian" site:	○	○	○	○	○

Figure 4.7: Third page of the Guardian survey



▼ DETAILS ON THE 20 MOST POPULAR TRACKERS

Doubleclick is an advertising company that develops and provides Internet ad serving services

Google is a searche engine for Internet

Google Analytics is a web analytics service offered by Google that tracks and reports website traffic,.

Facebook is a social media.

Google Adservices is a service that allows to insert advertising spaces within Google search pages.

Google Tag Manager is a tag management system (TMS) that allows you to quickly and easily update measurement codes and related code fragments collectively known as tags on your website or mobile app..

Amazon Adsystem associated with Amazon, allow you to create and monitor advertisements in an easy and intuitive way

Google Syndication is simply the Adsense advertisement network that displays Google ads on pages that contain these.

Rubicon Project is an expert society in digital advertising.

Yahoo is an American web services provider.

Pubmatic is a company that develops and implements online advertising software and strategies for the digital publishing and advertising industry

Scorecard Research Beacon a service of Full Circle Studies, Inc., is part of the Comscore, Inc. market research community, a leading global market research effort that studies and reports on Internet trends and behavior. ScorecardResearch conducts research by collecting Internet web browsing data and then uses that data to help show how people use the Internet, what they like about it, and what they don't.

ADform is a global digital media advertising technology company.

Bing is a Microsoft search engine .

Bidswitch is a middleware that allows connected programmatic partners to seamlessly access new platforms and services to optimizing their performance..

Bluekai is a cloud-based big data platform that enables companies to personalize online, offline, and mobile marketing campaigns.

Twitter is a social media.

Openx is a programmatic advertising technology company..

Zeotap is a Customer Intelligence (CIP) platform that helps companies better understand their customers and predict behaviors.

UnRuly Media is a leading video ad platform.

techniques: 'cookies'
type: advertising
Frequency: 25.00%

Figure 4.8: Details on top 20 trackers

4.2 Survey Dissemination

The purpose of this variety in the survey structure is therefore, as explained up to now, to try to involve the majority of the population and make them more sensitive to the issue of web tracking.

In order to reach this purpose, the last and crucial step is the survey dissemination through the media available to us. Social media, such as Instagram, Facebook and LinkedIn, have certainly turned out to be a powerful information mediums that used in the right way have allowed us to get our message across to many people quickly. In addition to using my personal profile on these platforms, the communication channels of the PimCity project were also used with the aim of receiving complete answers for the Spanish and English version of the survey.

However, they were not enough to arrive at a considerable number for a more precise analysis and, for this reason the survey was also sent, by e-mail, to students and teachers of the Polytechnic of Turin. The latter have shown a lot of interest in this topic and allowing us to reach a total of just over 1000 responses in just under a month, that can be considered enough for the purpose of this thesis work and to develop an automatic answers evaluation system.

Chapter 5

Analysis of survey answers

At this point, I started to analyze the 1038 answers gathered by the survey, thanks to the development of a Python code capable to derive statistical information useful for our purpose.

5.1 Preliminary section analysis

The first step of the analysis carried out consisted in the characterization of the interviewed population. To achieve this aim, the preliminary section answers of the survey have been analyzed, in which the interviewee was asked to answer to some personal questions and some theoretical questions. From these questions, therefore, it was possible to derive some percentages about the characteristics of users that have been plotted through pie charts.

The first plot regards the percentages of languages used by the users. This choice represents a crucial step of the survey because, on the basis of this choice, not only a language translation was done, but the version of the survey was determined. In fact, for each language chosen, a different dataset of 25 websites was taken and, for this reason, in the following steps it is needed to make a distinction between different versions.

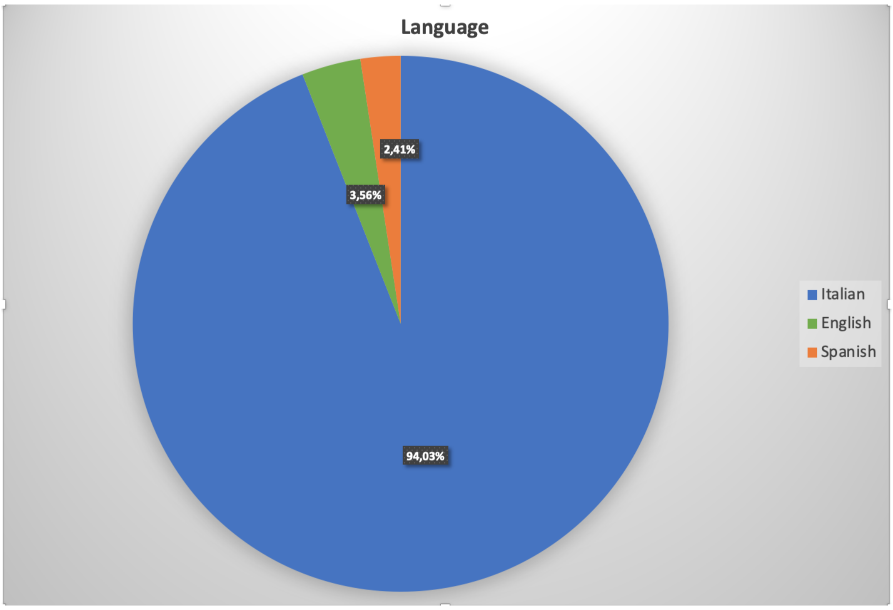


Figure 5.1: Pie chart languages used

In the pie chart, shown above fig.5.1, it is possible to note that, despite the attempts to involve people of different nationalities by disseminating the survey through different communication channels, 94% of the interviewees answered the questions of the Italian version.

However, the communication channels used had a big impact in the characterization of the interviewees, as showed in the following pie charts. In fact, social media and emails sent to students and teachers of the Polytechnic have led to have a fairly low average age of users with 53,47% of the interviewees with an age between 18 and 26 years old fig. 5.3 and the 54,05% of students, fig. 5.4. Moreover, almost the totality of the interviewees is well prepared with at least a high school diploma, as the pie chart shows in fig. 5.5: the 35,45% have an high school diploma, 22,74% have a bachelor's degree, 22,54% have a master's degree and the 17,44% have a PhD.

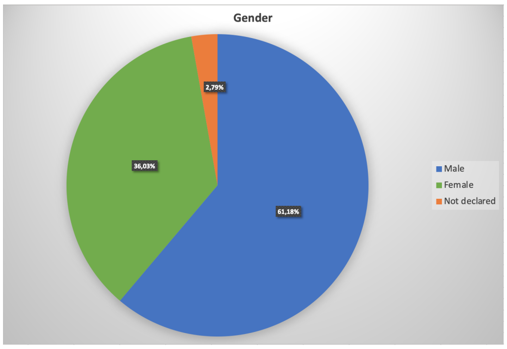


Figure 5.2: Pie chart users' gender

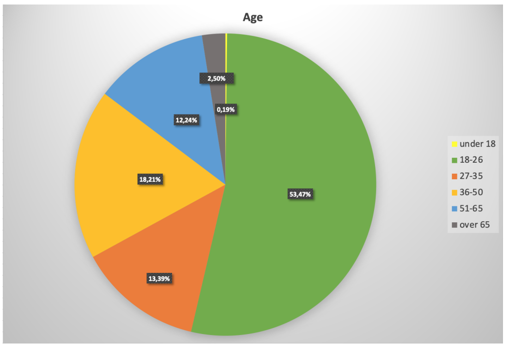


Figure 5.3: Pie chart users' age

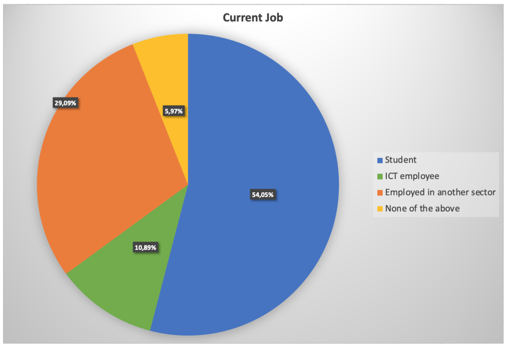


Figure 5.4: Pie chart users' current job

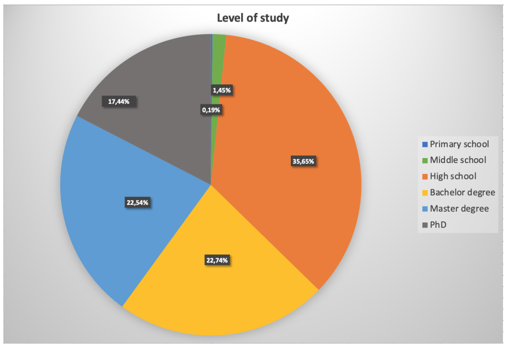


Figure 5.5: Pie chart users' level of study

This first characterization was confirmed by the subsequent analysis regarding the theoretical part. Indeed, the survey shows that over 50% of the interviewees answered correctly to the theoretical questions proposed, giving us the possibility to recognize a good level of the basic computer terminology of the interviewed population.

This level is recognized assigning a positive score for each correct answer and a negative score for a wrong one. In particular the valuation is done in the following way:

- for the multiple choice question: 1 point for every correct choice, -0.25 for every wrong choice;
- for the single choice question: 1 point for a correct answer, -0.5 for a wrong answer)

In this way the interviewee can achieve a total of 11 points. For this reason, three level have been recognized from the Python algorithm developed:

1. **expert level**, if the score achieved is between 8 and 11 points.
2. **medium level**, if the score achieved is between 4 and 8 points.
3. **beginner level**, if the score achieved is between 0 and 4 points.

In the figure below, fig. 5.6, it is therefore possible to see the pie chart relating to the percentages of the interviewed population levels where, as has been described so far, it is shown that most of the interviewees proved to have a good knowledge on the proposed topics and only 10.89 % of them did not reach the average level:

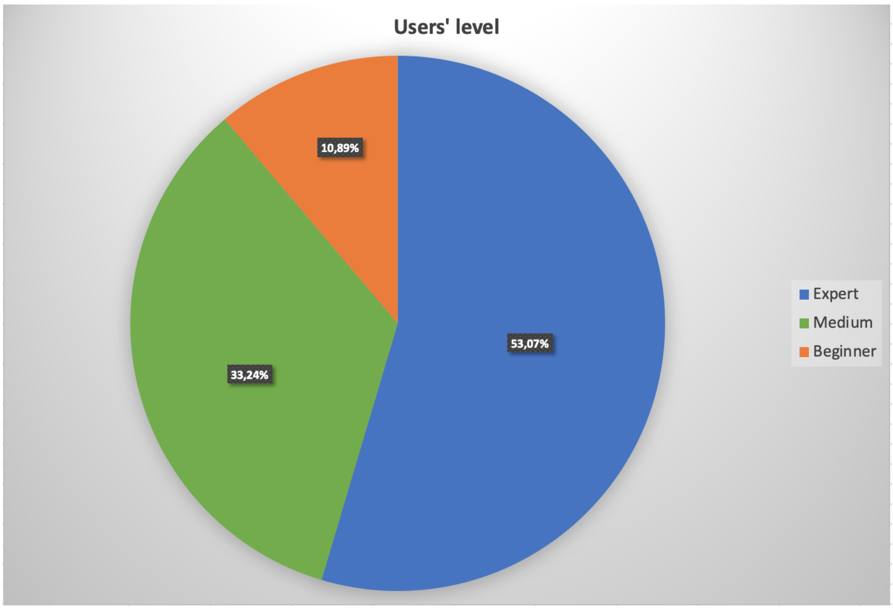


Figure 5.6: Pie chart users' level

5.2 Risk evaluation section analysis

Once the first part of the answers analysis was concluded, it was possible to analyze the behavior of the interviewees in the Risk evaluation section, analyzing how the different categories of users, previously described and analyzed, carried out the survey and what was their perceived risk for the websites presented. From the results obtained in the previous section and since different version of the survey are available, it is necessary to specify that the analysis conducted in the following are only related to the italian version of the survey.

As explained in the previous chapter, this second section consists of three pages and, for this reason, it is significant to analyze not only the perceived risk on the last page, but also the previous reactions on the first two pages that then led to the final evaluation of the website. In fact, as it is possible to see in fig. 4.5 and 4.6, the interviewee was asked to make two preliminary assessments of the website in question and these steps can be fundamental for understanding change in the user's perception while the data trackers are slowly presented to him.

For this purpose, the first analysis conducted is, therefore, related to this topic. In fact, all the answers to the questions on the first and second page have been gathered and used to make a scatter plot that illustrates, as the number of trackers present on the website increases, the change in the user's reaction before and after this kind of data on web tracking was presented to him. This change in the users' reactions have been computed considering the percentage increase of the score obtained in the first and the second page of the survey (in a scale from 1 -> I trust, to 3 -> I do not trust). The idea is to underline the difference (if there is) between the first user impression in the first page, where the user was unaware of the data trackers behind the website presented, and the user reaction in the second page, where the data trackers are revealed to the user.

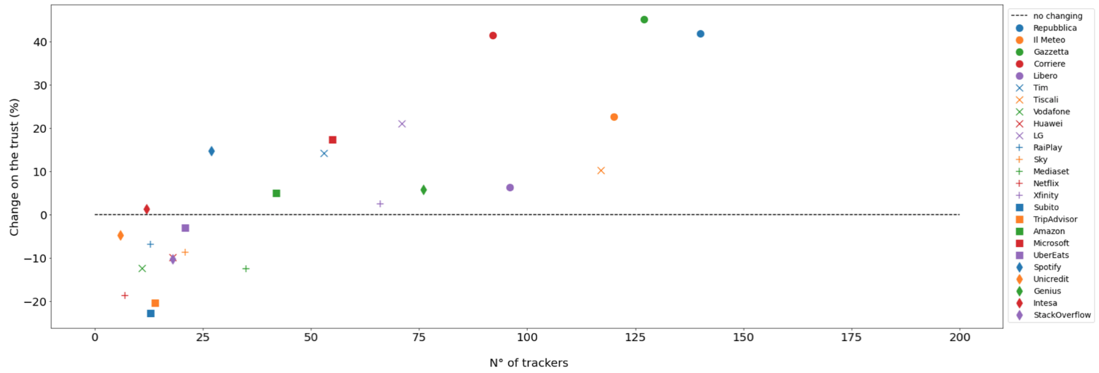


Figure 5.7: Scatter plot Change on the trust (%) vs number of trackers per websites

The scatter plot shown above represents, therefore, in the x axis, the number of trackers in ascending order and, in the y axis, the change in the trust in percentage. For every category of websites different symbols have been chosen and for every websites belonging to each category five different colours have been chosen in order to make a visible distinction between the different categories and websites examined. As it is possible to see the increase of the number of trackers, that are shown in the second page of the survey, had a big impact in the user's reaction. In fact, for almost every websites that have less than 25 trackers the obtained percentage is negative and this means that the average first impression of the users is worse than the successive reaction in the second page of the survey. Instead, with a number of trackers shown above 25, it is possible to notice how the users interviewed showed a greater concern for the security of their data, showing a percentage increase in the perception of risk from the first impression just by looking at the home page and some general features to the first reaction to the data provided on the trackers. This impact was more evident for the newspaper category where, in particular for the websites of the "Gazzetta", "La Repubblica" and "Il Corriere", the percentage increase in perceived risk on average between the first impression and the subsequent reaction exceeded 40 % .This means that users had a good first impression for these category of sites that completely changes once they look at their web tracking data, showing a net drop in trust as a result of an increase in the computed percentage.

After the first two evaluations, the interviewee, looking also to the characteristics of the trackers that can be encountered in the website fig. 4.7, is asked to evaluate, in a scale from 1 to 5 (1 -> my data are not at risk, 5 -> my data are at risk), the overall perceived risk. These evaluations have been summarized in the following through an histogram plot that can give us a first idea of the final users perceptions on the websites analyzed in the survey:

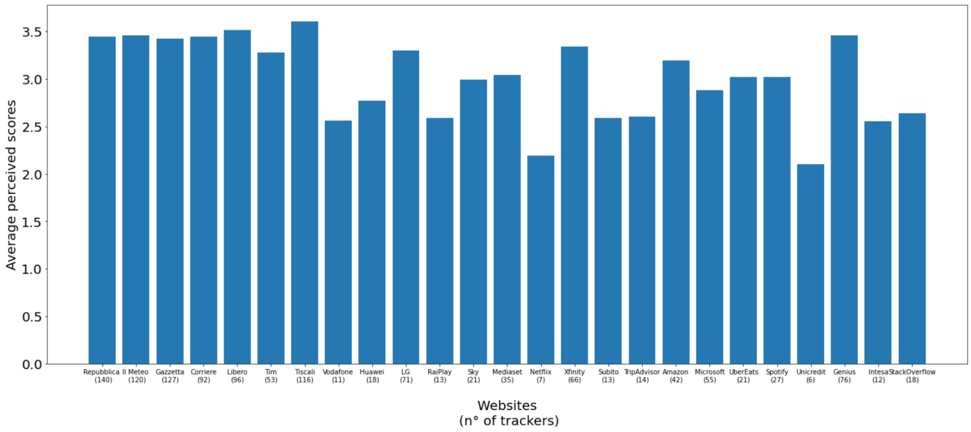


Figure 5.8: Average perceived scores per websites

In the figure above it is possible to see all the average perceived risk scores, in the y-axis, obtained for each websites of our dataset, in the x-axis. For each website, it is possible to immediately see the corresponding number of trackers to have a more clear idea of the histogram trend. The first category, the newspaper one, as anticipated in the previous scatter plot, obtained the greater scores, with scores of about 3.5 for all 5 websites, but the greater one was obtained by "Tiscali" with 3.6. The latter has not the biggest number of tracker and this confirmed us that not only this parameter affects the user perception but also other factors that will be analyzed forward. Instead, for what concern the smallest score obtained, it is possible to notice that this result is achieved by "Unicredit" website with a score of 2.1, following by "Netflix" with 2.19.

5.3 Statistical Analysis

After analyzing and discussing both the results of the preliminary section and the results of the risk evaluation section obtained from the survey conducted, the next step was to analyze the combination of the two sections. The idea was to carry out statistical evaluations of the answers obtained in the final part of the survey taking into account their personal information that was provided at the beginning of the survey in order to deepen the answers provided by the interviewees and understand which of the categories extrapolated from the preliminary section have had a major influence on the average perceived risk. For this purpose, the perceived risk scores on the last page of the risk evaluation section were gathered for each of the following groups:

- **Age category**
  1. 18-26 age group
  2. 26-35 age group
  3. 36-50 age group
  4. 51-65 age group
- **Gender category**
  1. Male group
  2. Female group
- **Job category**
  1. Student group
  2. ICT employee group

3. Employee in other sector group
4. None of the above group

- **School level category**

  1. High school diploma group
  2. Bachelor degree group
  3. Master degree group
  4. PhD group

- **User level category**

  1. Expert user group
  2. Medium user group
  3. Beginner user group

Statistical test is a tool that allow to quantify, within observed data, behaviours that would distinguish the null from the alternative hypothesis. The idea is, therefore, to start from a null hypothesis and with a statistical test understand if confirm or reject the hypothesis. Our first question to solve is to understand if the data obtained for the different categories of interviewees selected are normal distributed. For this purpose, the Anderson-Darling Test was chosen, that is a statistical test of whether a given sample of data is drawn from a given probability distribution. In its basic form, the test assumes that there are no parameters to be estimated in the distribution being tested, in which case the test and its set of critical values is distribution-free. However, the test is most often used in contexts where a family of distributions is being tested, in which case the parameters of that family need to be estimated and account must be taken of this in adjusting either the test-statistic or its critical values. When applied to testing whether a normal distribution adequately describes a set of data, it is one of the most powerful statistical tools for detecting most departures from normality. The results obtained from the first question are all the same, in fact, all the data distribution are not normal distributed. For this reason, to understand if the categories analyzed can be considered statistically different from each other, we cannot use a parametric tests, such as T-test, but we have to use a non parametric one. To achieve our aim, we used the Mann-Whitney test, (also known as the Mann-Whitney U test), that is one of the most powerful non-parametric tests for checking whether two statistical samples come from the same population. The null hypothesis, for this kind of test, is that the distributions of both groups are identical and in order to understand the result of the test we have to look on the p-value. The Mann-Whitney test first ranks all the values from low to high, and then computes a P value that depends on the discrepancy between the mean ranks of the two groups. Looking at the p-value two possible observations can be done:

- if the P value is small, under the 5 %, it is possible to reject the null hypothesis and conclude that the populations are distinct.
- if the P value is large, above the 5 %, the data do not give you any reason to reject the null hypothesis. This is not the same as saying that the two populations are the same, but you just have no compelling evidence that they differ.

5.3.1 Age category

The first category considered is the Age one for which the statistical tests, described before, were computed, comparing the distributions of the perceived risk scores and obtaining the following results:

Combination of age groups	**p-value**
(18-26)-(27-35) groups	$2.31 * 10^{-3}$
(18-26)-(36-50) groups	$4.67 * 10^{-6}$
(18-26)-(51-65) groups	$5.05 * 10^{-10}$
(27-35)-(36-50) groups	$1.83 * 10^{-1}$
(27-35)-(51-65) groups	$4.56 * 10^{-3}$
(36-50)-(51-65) groups	$2.89 * 10^{-2}$

Table 5.1: Mann-Whitney test for age groups

From the table above it is possible to see that for all the combinations of age groups the null hypothesis can be rejected apart from the combination (27-35)-(36-50) where the p-value exceed the 5% and it is possible to consider that there is no a statistical difference between the two distributions. For the other combinations considered, instead, we can say that there is a statistical difference and this result can give us a hint for the successive analysis. In fact, the Mann-Whitney tests allows us to think that the interviewees, coming from these categories, had a different approach to the survey giving different answers. To deepen this consideration, a density histogram plot was built which aims to show what were the differences in the answers given by the categories analyzed:

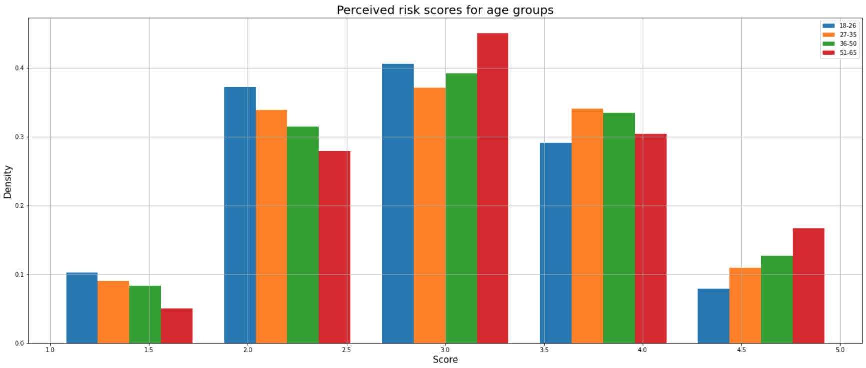


Figure 5.9: Density plot for the age category

As anticipated before, from the graph above, it is possible to notice, in particular the different trend between the distributions of the two age groups at the extremes, the 18-26 age group and 51-65 age group. In fact, the perceived risk of the younger group (blue block) is lower than the older one (red block) which instead shows peaks for higher scores, showing a greater perception of the risk for the security of their data while browsing on the internet.

5.3.2 Gender category

The same process done for the Age groups was repeated for the Gender category. The Mann-Whitney test done for the two distribution male and female results into a statistical difference with a p-value equal to 0.0289, giving us the possibility to reject the null hypothesis. For this reason, a density plot, was built also in this case, obtaining the following graph:

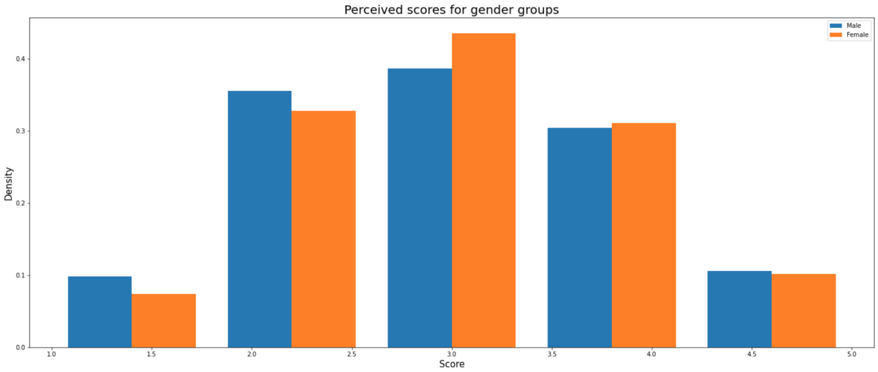


Figure 5.10: Density plot for the gender category

In this case, we can underline only a little difference in the first three scores, where the male group shows a lower perceived risk than the female group showing peaks in the 1 and 2 score, instead the female group have a peak for the 3 score.

5.3.3 Job category

As in the previous cases, starting from H0, null hypothesis (two groups have the same distributions) we computed the Mann-Whitney test and we built the following table:

Combination of job groups	**p-value**
(student)-(ICT employee) groups	$1.43 * 10^{-3}$
(student)-(Employee in other sector) groups	$6.40 * 10^{-8}$
(student)-(None of the above) groups	$2.34 * 10^{-2}$
(ICT employee)-(Employee in other sector) groups	$3.61 * 10^{-1}$
(ICT employee)-(None of the above) groups	$3.64 * 10^{-1}$
(Employee in other sector)-(None of the above) groups	$2.44 * 10^{-1}$

Table 5.2: Mann-Whitney test for job groups

In this case, we have to reject H0 for the first three combinations of the table and it is possible to immediately notice that all the three lines regards the student group. For this reason, also in this case, a density histogram plot was built in order to understand better what the Mann-Whitney test try to evidence.

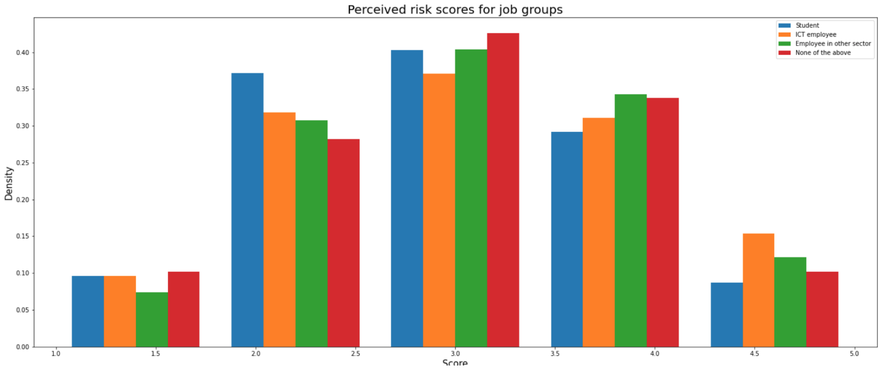


Figure 5.11: Density plot for the job category

As it possible to see from the plot, students have a perceived risk lower than other groups having peaks in the left side of the density plot and lower peaks in the right side, confirming the difference anticipated from the Mann-Whitney test.

5.3.4 School category

Similar results to the previous case have been obtained for the school category in the Mann-Whitney test. In fact, from the p-value table obtained, shown below 5.3, it is possible to notice that only the first three combinations, regarding in this case the high school diploma group, allows us to reject the null hypothesis.

Combination of school groups	**p-value**
(high school diploma)-(Bachelor degree) groups	$4.63 * 10^{-8}$
(high school diploma)-(Master degree) groups	$3.63 * 10^{-5}$
(high school diploma)-(phD) groups	$5.49 * 10^{-5}$
(Bachelor degree)-(Master degree) groups	$1.00 * 10^{-1}$
(Bachelor degree)-(phD) groups	$1.55 * 10^{-1}$
(Master degree)-(phD) groups	$4.31 * 10^{-1}$

Table 5.3: Mann-Whitney test for school groups

Moreover it is shown below the density histogram graph for school categories where it is possible to see how high school graduates showed higher perceived risk, determined by a significant peak in score 5, demonstrating that they had more than once the feeling that their data was at high risk on the websites presented to them.

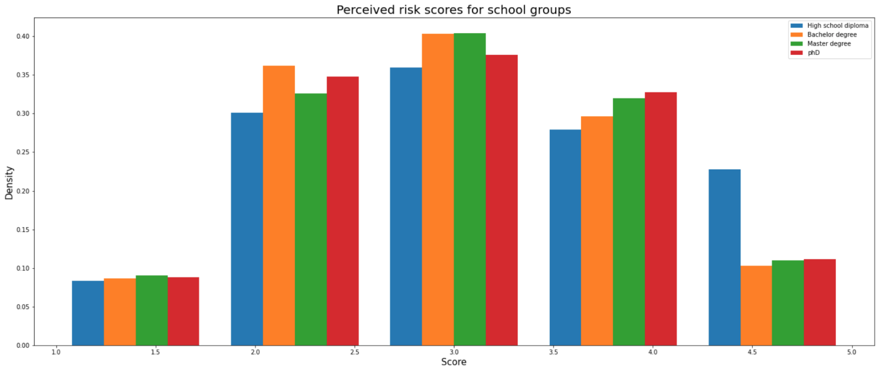


Figure 5.12: Density plot for the school category

5.3.5 User level category

The last statistical analysis regards the user level determined in the theory section. Also in this case, the Mann-Whitney test was computed for all the three possible combinations obtaining the following table:

Combination of user level groups	**p-value**
(expert level)-(beginner level) groups	$2.41 * 10^{-7}$
(expert level)-(medium level) groups	$5.26 * 10^{-9}$
(beginner level)-(medium level) groups	$1.58 * 10^{-1}$

Table 5.4: Mann-Whitney test for school groups

In this case the expert level distribution results statistical different from the other two distribution and it is confirmed from the density plot in the following where it is possible to notice immediately how the expert group have never choose the 5 score during their survey. They, in fact, show a major density for the 2 and 3 scores demonstrating a lower perceived risk than other two groups.

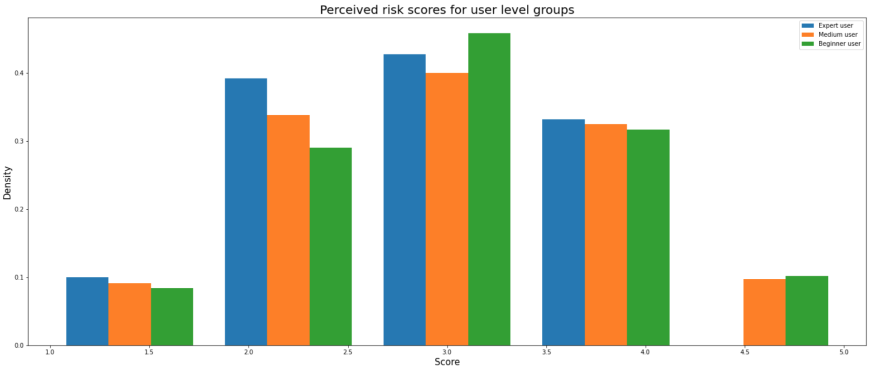


Figure 5.13: Density plot for the user level category

Chapter 6

Prediction model of the privacy score

6.1 Regression algorithm

The final goal of this thesis work, as mentioned in the paragraph 1.2, was to use machine learning techniques in order to develop a model to support the construction of the Transparency Tag (TT). Machine learning is the study of computer algorithms that can improve automatically through experience and by the use of data. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. In particular, in this thesis work, the focus was on regression analysis, one of the most basic tools in the area of machine learning used for prediction. It includes a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome' or 'response' variable) and one or more independent variables (often called 'predictors'). The most common form of regression analysis is linear regression, which assumed that the relationship between the dependent variable y and the independent variables X=[x1,x2...] is linear. This relationship is modeled through an error variable ϵ, an unobserved random variable that adds "noise" to the linear relationship between the dependent variable and input features. Thus the model takes the form:

$$Y = \beta_0 + \beta_i x_i + \epsilon \tag{6.1}$$

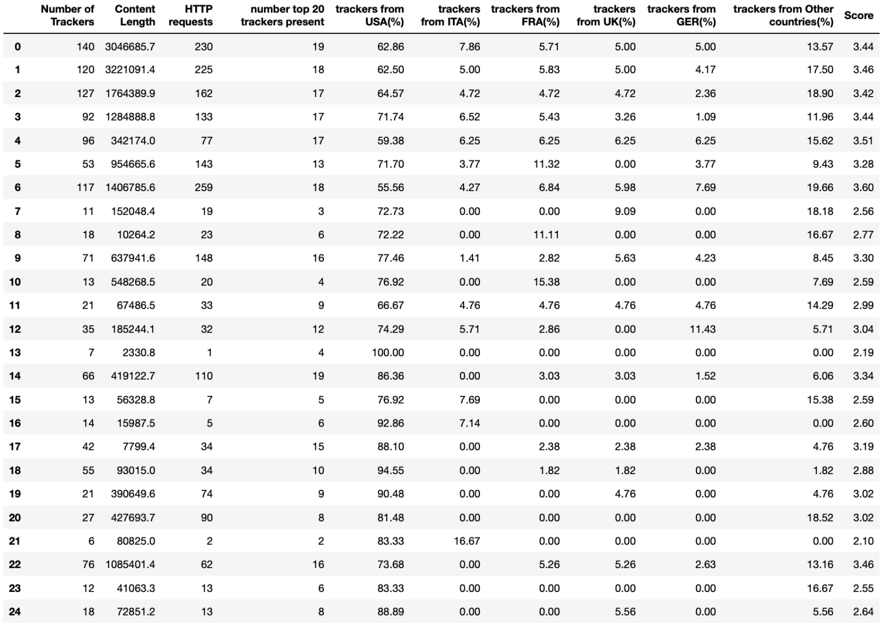
where β is the intercept term and β_i are the regression coefficients.

6.2 Model construction

The idea was, therefore, to apply linear regression to the answers gathered from the survey to reach our final goal, exploiting this kind of algorithm, implemented by the scikit-learn library in Python [13]. The average perceived risk, obtained on the last page of the survey, was identified as dependent variable Y and objective data, collected from the dataset provided by Ermes described in chapter 3, were identified as independent variables x_i. In order to understand better the algorithm and the impact of the input features on dependent variable more than one training dataset were built where the number of independent variables used changes. For the regression model, therefore, different dataframes (two-dimensional data structures with labeled axes) have been created on python, with a number of rows equal to the number of website analyzed (25) and a number of columns equal to the number of input features selected plus the variable dependent. The input features that have been used are the same of the objective data presented to the interviewees during the survey, in particular in the last two pages for every website:

- Number of trackers (linear or logarithmic);
- Content length (linear or logarithmic);
- HTTP requests (linear or logarithmic);
- the number of top 20 trackers presented in the website (linear or logarithmic);
- percentages of the website trackers on the basis of their country origin;
- the type of website;

Therefore, 6 different dataframes have been built, starting from the first dataframe with only three input features of them as independent variables and finishing with all these objective data selected (an example is shown in figure 6.1).



	Number of Trackers	Content Length	HTTP requests	number top 20 trackers present	trackers from USA(%)	trackers from ITA(%)	trackers from FRA(%)	trackers from UK(%)	trackers from GER(%)	trackers from Other countries(%)	Score
0	140	3046685.7	230	19	62.86	7.86	5.71	5.00	5.00	13.57	3.44
1	120	3221091.4	225	18	62.50	5.00	5.83	5.00	4.17	17.50	3.46
2	127	1764389.9	162	17	64.57	4.72	4.72	4.72	2.36	18.90	3.42
3	92	1284888.8	133	17	71.74	6.52	5.43	3.26	1.09	11.96	3.44
4	96	342174.0	77	17	59.38	6.25	6.25	6.25	6.25	15.62	3.51
5	53	954665.6	143	13	71.70	3.77	11.32	0.00	3.77	9.43	3.28
6	117	1406785.6	259	18	55.56	4.27	6.84	5.98	7.69	19.66	3.60
7	11	152048.4	19	3	72.73	0.00	0.00	9.09	0.00	18.18	2.56
8	18	10264.2	23	6	72.22	0.00	11.11	0.00	0.00	16.67	2.77
9	71	637941.6	148	16	77.46	1.41	2.82	5.63	4.23	8.45	3.30
10	13	548268.5	20	4	76.92	0.00	15.38	0.00	0.00	7.69	2.59
11	21	67486.5	33	9	66.67	4.76	4.76	4.76	4.76	14.29	2.99
12	35	185244.1	32	12	74.29	5.71	2.86	0.00	11.43	5.71	3.04
13	7	2330.8	1	4	100.00	0.00	0.00	0.00	0.00	0.00	2.19
14	66	419122.7	110	19	86.36	0.00	3.03	3.03	1.52	6.06	3.34
15	13	56328.8	7	5	76.92	7.69	0.00	0.00	0.00	15.38	2.59
16	14	15987.5	5	6	92.86	7.14	0.00	0.00	0.00	0.00	2.60
17	42	7799.4	34	15	88.10	0.00	2.38	2.38	2.38	4.76	3.19
18	55	93015.0	34	10	94.55	0.00	1.82	1.82	0.00	1.82	2.88
19	21	390649.6	74	9	90.48	0.00	0.00	4.76	0.00	4.76	3.02
20	27	427693.7	90	8	81.48	0.00	0.00	0.00	0.00	18.52	3.02
21	6	80825.0	2	2	83.33	16.67	0.00	0.00	0.00	0.00	2.10
22	76	1085401.4	62	16	73.68	0.00	5.26	5.26	2.63	13.16	3.46
23	12	41063.3	13	6	83.33	0.00	0.00	0.00	0.00	16.67	2.55
24	18	72851.2	13	8	88.89	0.00	0.00	5.56	0.00	5.56	2.64

Figure 6.1: Example of dataframe used

Once the training datasets were constructed, it was developed the python code for the regression algorithm in order to understand the performance reached by each model. To estimate the performance of linear regression algorithm the Leave-One-Out Cross-Validation (LOOCV) was used. With this procedure each sample of the dataset is used once as a test set (singleton) while the remaining samples form the training set, creating a number of test sets equal the number of samples. LOOCV provides therefore reliable and unbiased estimate of model performance, but due to the high number of test sets this cross-validation method can be a computationally expensive procedure to perform. For this reason, LOOCV is appropriate with small dataset (as in this case) or when an accurate estimate of model performance is more important than the computational cost of the method. In the following is shown the code developed for the LOOCV procedure described:

regression_model.py

```
#!/usr/bin/env python
# coding: utf-8

# In[ ]:


# importing train_test_split from sklearn
from sklearn.model_selection import train_test_split
# importing module
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import LeaveOneOut
from sklearn.metrics import accuracy_score
from sklearn.model_selection import cross_val_score
from numpy import mean
from numpy import absolute
loo = LeaveOneOut()
#selecting from the dataframe only the independent variables into x
x = df_reg.drop('Score',axis=1)
#separate the predicting attribute into Y for model training
y = df_reg['Score']
# splitting the data to obtain training data and test data
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2, random_state = 42)
# creating an object of LinearRegression class
LR = LinearRegression()
# fitting the training data
model = LR.fit(x_train,y_train)
# evaluate model
scores = cross_val_score(model, x, y, scoring='neg_root_mean_squared_error',cv=loo, n_jobs=-1)
# report performance with RMSE
print("RMSE:",mean(absolute(scores)))


# In[ ]:


# R2 score evaluation
ytests = []
ypreds = []
for train_idx,test_idx in loo.split(x):
    X_train, X_test = x.iloc[train_idx], x.iloc[test_idx] #requires arrays
    y_train, y_test = y.iloc[train_idx], y.iloc[test_idx]

    model = LinearRegression()
    model.fit(X = X_train, y = y_train)
    y_pred = model.predict(X_test)

    # there is only one y-test and y-pred per iteration over the loo.split ,
    # so to get a proper graph, we append them to respective lists.
    ytests += list(y_test)
    ypreds += list(y_pred)

rr = r2_score(ytests, ypreds)

print("Leave One Out Cross Validation")
print("R2 score: {:.2f}%".format(rr*100))
```

6.3 Results obtained from the model

As shown above, from the code used to develop the linear regression algorithm, it is possible to notice that to evaluate the performance of each model developed the root mean squared error (RMSE) and R-squared (R2) score have been used. RMSE and the R2 are metrics used to measure the accuracy reached by a linear regression model and to assess how well it fits a dataset. The RMSE is the square root of the variance of the residuals and, therefore, corresponds to the standard deviation of the residuals (prediction errors), representing a measure of how far apart the predicted values are from the observed values in a dataset, on average. R2 score represents, instead, the proportion of the variation in the dependent variable that is predictable from the independent variable, telling us how well the predictor variables can explain the variation in the response variable. The two metrics have therefore the same purpose but are different and in order to have a complete view of the model performance in this thesis work both have been computed. Below are shown the results obtained from the model constructed with the data obtained from the survey.

Features used	**dat 1**	**dat 2**	**dat 3**	**dat 4**
N° of trackers	Y-Lin	Y-Lin	Y-Lin	Y-Lin
Content Length	Y-Lin	Y-Lin	Y-Lin	Y-Lin
HTTP requests	Y-Lin	Y-Lin	Y-Lin	Y-Lin
N° of top 20 trackers present	N	Y-Lin	Y-Lin	Y-Lin
% country origin	N	N	Y	Y
type of website	N	N	N	Y
RMSE score	0.19	0.13	0.13	0.12
R2 score	68.66%	85.31%	85.91%	90.57%

Table 6.1: Results of Linear Regression with linear expression

Features used	dat 5	dat 6	dat 7	dat 8
N° of trackers	Y-Log	Y-Log	Y-Log	Y-Log
Content Length	Y-Log	Y-Log	Y-Log	Y-Log
HTTP requests	Y-Log	Y-Log	Y-Log	Y-Log
N° of top 20 trackers present	N	Y-Log	Y-Log	Y-Log
% country origin	N	N	Y	Y
type of website	N	N	N	Y
RMSE score	0.09	0.08	0.07	0.1
R2 score	93.46%	94.61%	95.20%	92.70%

Table 6.2: Results of Linear Regression with logarithmic expression

In the tables above are therefore shown all the results obtained from the regression algorithm applied on the different dataframes built (called "dat" in the tables). From the tables it is possible to understand how datasets are built seeing which features are used (Y), in which numeric expression (Linear or Logarithmic) and the accuracy reached in terms of RMSE and R2 score in the last two rows. RMSE is always non-negative, and a value of 0 (almost never achieved in practice) would indicate a perfect fit to the data. In general, a lower RMSE is therefore better than a higher one. R2 score, instead, measures the strength of the relationship between the model and the dependent variable on a convenient 0 – 100% scale. The two tables differs only on the the numeric expression of the first four features (N° of tracker, content length, HTTP requests and number of top 20 trackers present) that are represented in the linear way in the table 6.1 and in the logarithmic way in the table 6.2. The structure of the two tables is, instead, the same. In fact, starting from the first column to the last a new feature is added on the datasets in order to understand which of them have a positive or negative impact on the accuracy reached. For what concern the first table 6.1 (with linear parameters) it is possible to see how, for all the four dataframes, the model reach a good level of accuracy. It easy to notice how each added feature provides a positive contribution to the model, improving the model prediction performance up to achieve an R2 score of 90.57% and RMSE score of 0.12 with the complete dataset dat 4. In particular, it is also possible to notice how the addition of number of top 20 tracker to the dataset 1 had the greatest impact, passing from 68.66% of accuracy in dat 1 to 85.31% in dat 2. Looking, instead, on the second table 6.2, it is possible to notice how better performance than before have been reached. From these results it is clear that the logarithmic expression have a big impact to reduce the RMSE and improve the accuracy in prediction. Also in this case, the features addition improved the model performance but up to dat 7. In fact, with the addition of the category feature in dat 8 the performance suffer a small decline. The best result obtained on all dataframes proved was therefore the one reached with dat 7, the dataset with all the first four features expressed in the logarithmic way and the country origin feature, that achieve a root mean square error (RMSE) of 0.07 and an R-squared of 95.20%. To show the importance of the results of this model a scatter plot has been built with the prediction values on y-axis and the test valus on x-axis:

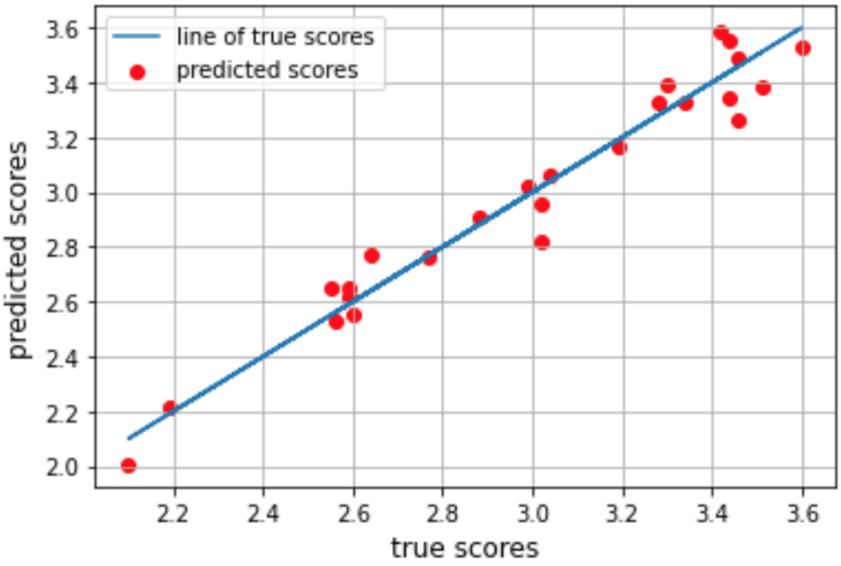


Figure 6.2: Scatter plot of the best prediction model

The figure above shows how the predicted perceived privacy scores (red points) fits very well the line of the true values (blue line) and this confirms the RSME and R2-score results that provide us a proof of how the model constructed is a very good prediction model.

To better understand the impact of the features on the risk perceived scores obtained a further analysis was conducted. In fact, it was exploited the Pearson correlation to measure the linear correlation of two sets of data, the independent data x_i and the dependent data Y. The Pearson correlation is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between -1 and 1. To reach our purpose a correlation measure was done between each feature x_i used and the score Y.

Features	Pearson correlation
Number of Trackers	0.96
number top 20 trackers present	0.95
HTTP requests	0.93
Content Length	0.72
trackers from GER(%)	0.59
Type:newspaper	0.54
trackers from UK(%)	0.49
trackers from Other countries(%)	0.45
trackers from FRA(%)	0.37
Type:music	0.17
Type:electronics	0.12
trackers from ITA(%)	-0.07
Type:Shopping online	-0.17
Type:other categories	-0.173113
Type:Tv movies and streaming	-0.20
Type:banks	-0.47
trackers from USA(%)	-0.63

Table 6.3: Pearson correlation between features used and score

From the table above, it is easy to notice how the impact of the features regarding the number of trackers, the number of top 20 trackers and the number of HTTP requests had a big impact on the score with a very high correlation, demonstrating the very good accuracy resulted for the dataset 7. It is also important to notice the very low correlation between the fraction of trackers that coming from the USA and the score, in fact all the websites analyzed presents an high percentage and for this reason it is not a characteristic relevant for the linear regression model where important features are instead those that underline the difference between different websites and as a consequence different scores.

Chapter 7

Conclusion and Future works

The results obtained during this thesis work show, therefore, the world of web tracking observed from a different point of view, that relating to the user side. The web tracking threat has been analyzed from an objective point of view, in most of the cases, observing the evolution of the tracking technologies used and their pervasiveness, indirectly calculating the risk that any user, who browses the web, could perceive from it. With this work, on the other hand, it was possible to know directly the level of risk perceived by users, bringing them closer to this topic thanks to the survey conducted and then analyzing the results obtained using the model developed. With the latter it is possible to understand which information the user is most sensitive to and predict the subjective risk perceived by the user for all websites for which some specified objective characteristics are known, as described in the previous chapter. This work can certainly be subject to further improvements by using other techniques or by making other statistical evaluations on the answers collected by the survey, obtaining more information than that analyzed during this process, but it can certainly provide the basis for future work in this way.. For this reason, the model developed can become an important starting tool for the PimCity project. In fact, it can help to build detailed TTs (Transparency Tags) for websites and used along with a PDA (Personal Data Avatar), they will improve users awareness of the web tracking ecosystem and help protect data privacy.

List of Tables

List of Figures

Acknowledgements

I want to thank the supervisors of this thesis, Prof. Marco Mellia, Dott. Martino Trevisan, and Dott. Luca Vassio, who have guided me during the work in the best way possible.

This work marks the end of a long journey full of ups and downs faced thanks to the support of my family and friends to which I want to dedicate an immense "THANK YOU" here too.

Bibliography

[1] Wikipedia. *General Data Protection Regulation - Wikipedia.* URL: `https://en.wikipedia.org/wiki/General_Data_Protection_Regulation` (cit. on p. 2).

[2] *PimCity Project.* URL: `https://pimcity.com/` (cit. on p. 2).

[3] *Ermes - Web Protection and cyber security.* URL: `https://ermes.company/it/` (cit. on p. 3).

[4] Adam Lerner, Anna Kornfeld Simpson, Tadayoshi Kohno, and Franziska Roesner. «Internet jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016». In: *25th {USENIX} Security Symposium ({USENIX} Security 16).* 2016 (cit. on p. 4).

[5] Balachander Krishnamurthy and Craig Wills. «Privacy diffusion on the web: a longitudinal perspective». In: *Proceedings of the 18th international conference on World wide web.* 2009, pp. 541–550 (cit. on p. 4).

[6] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. «The web never forgets: Persistent tracking mechanisms in the wild». In: *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security.* 2014, pp. 674–689 (cit. on p. 4).

[7] Steven Englehardt and Arvind Narayanan. «Online tracking: A 1-million-site measurement and analysis». In: *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security.* 2016, pp. 1388–1401 (cit. on p. 5).

[8] Martino Trevisan, Stefano Traverso, Eleonora Bassi, and Marco Mellia. «4 Years of EU Cookie Law: Results and Lessons Learned.» In: *Proc. Priv. Enhancing Technol.* 2019.2 (2019), pp. 126–145 (cit. on p. 5).

[9] William Melicher, Mahmood Sharif, Joshua Tan, Lujo Bauer, Mihai Christodorescu, and Pedro Giovanni Leon. «Preferences for Web Tracking». In: *Proceedings on Privacy Enhancing Technologies* 2016.2 (2016), pp. 1–20 (cit. on p. 5).

[10] Jaspreet Bhatia and Travis D Breaux. «Empirical measurement of perceived privacy risk». In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 25.6 (2018), pp. 1–47 (cit. on p. 5).

[11] Valentino Rizzo, Stefano Traverso, and Marco Mellia. «Unveiling Web Fingerprinting in the Wild Via Code Mining and Machine Learning.» In: *Proc. Priv. Enhancing Technol.* 2021.1 (2021), pp. 43–63 (cit. on p. 5).

[12] *Tranco ranking.* URL: `https://tranco-list.eu` (cit. on p. 7).

[13] *Scikit-learn documentation.* URL: `https://scikit-learn.org/0.21/documentation.html` (cit. on p. 36).