Instituto Superior
de Contabilidade
e Administração
Politécnico de Coimbra

ISCAC | 2023 Vitor Monteiro Pinto | Generating Machine Learning-led Insights for Customer-Centric Web Design and Audience Strategy: The case of a Portuguese higher education institution website


Instituto Superior
de Contabilidade
e Administração
Politécnico de Coimbra
COIMBRA BUSINESS SCHOOL
ISCAC.pt



Vitor Monteiro Pinto


# Generating Machine Learning-led Insights for Customer-Centric Web Design and Audience Strategy: The case of a Portuguese higher education institution website


Coimbra, outubro de 2023

Instituto Superior de Contabilidade e Administração
Politécnico de Coimbra

COIMBRA BUSINESS SCHOOL
ISCAC.pt

Vitor Monteiro Pinto

# Generating Machine Learning-led Insights for Customer-Centric Web Design and Audience Strategy: The case of a Portuguese Higher education institution website

Trabalho de projeto submetido ao Instituto Superior de Contabilidade e Administração de Coimbra para cumprimento dos requisitos necessários à obtenção do grau de **Mestre em Análise de Dados e Sistemas de Apoio à Decisão,** realizado sob a orientação dos Professores Fernando Paulo Belfo e Isabel Pedrosa e supervisão da Professora Ana Ferreira.

Coimbra, outubro de 2023

## TERMO DE RESPONSABILIDADE

Declaro ser a autor(a) desta dissertação / projeto / relatório de estágio, que constitui um trabalho original e inédito, que nunca foi submetido a outra Instituição de ensino superior para obtenção de um grau académico ou outra habilitação. Atesto ainda que todas as citações estão devidamente identificadas e que tenho consciência de que o plágio constitui uma grave falta de ética, que poderá resultar na anulação da(o) presente dissertação / projeto / relatório de estágio.


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

# DEDICATÓRIA

Dedico à minha família, que apoiou constantemente a minha ida para outro continente para os traçar novos caminhos acadêmicos e pessoais.

Dedico aos amigos internacionais que fiz durante a minha estadia em Coimbra e que levo até hoje e que foram uma base essencial de acolhimento, de parceria e de bom ambiente para alcançar resultados de sucesso em todos os campos da vida.

Dedico à Coimbra, a cidade dos estudantes, por me acolher como uma casa de transformação e novos começos. A famosa frase "Uma vez Coimbra, para sempre saudade" fica como uma verdade na minha memória.

# AGRADECIMENTOS

Agradeço aos docentes do mestrado, que me deram uma nova visão sobre os dados em suas aplicações, técnicas que foram de grande importância para o meu desenvolvimento intelectual e profissional.

Agradeço aos meus orientadores, os professores doutores Isabel Pedrosa e Fernando Paulo Belfo, que acreditaram no potencial do projeto, que foram essenciais em seu desenvolvimento técnico, mas principalmente no seu apoio em unir pessoas e organizações para torná-lo possível em prol de abrir caminho de novas oportunidades e de resultados para todos os envolvidos.

Agradeço à equipe do Instituto Politécnico de Coimbra, nomeadamente a Vice-Presidente Ana Ferreira, a chefe de divisão do Gabinete de Comunicação Institucional e Imagem (GCII) e os representantes Jéssica Lopes da Comunicação Institucional e Jorge Pardal do Departamento de Tecnologias de Informação e Comunicação (DTIC), que enxergaram o valor do projeto e que dedicaram tempo para sua realização pela experiência como profissionais e como novas formas de otimizar resultado para a instituição nos meios digitais.

Agradeço também à equipe do Instituto Superior de Contabilidade e Administração de Coimbra, que deu apoio individualizado de alto nível para os meus projetos pessoais e profissionais que significaram um grande diferencial para a minha experiência no período como estudante e para meu futuro laboral.

# RESUMO

Ao buscar uma instituição de ensino superior (IES) para matricular-se, os futuros alunos têm contato com diversos canais dessas instituições ao longo de sua jornada. O website institucional está entre esses canais e a forma como ele é projetado pode influenciar em quão engajados os seus visitantes ficam. O comportamento dos usuários é monitorado por meio da coleta de dados de suas interações em ferramentas de Web Analytics, permitindo a sua análise e, consequentemente, a geração de insights, que podem ser utilizados na melhora do engajamento dos usuários no site. Esse estudo de caso propõe o uso de técnicas de Data Mining para a geração de insights em uma IES portuguesa, com apoio do método CRISP-DM. Para poder coletar, armazenar e transformar os dados foram utilizadas as ferramentas Google Tag Manager, Analytics, BigQuery e Data Studio. Já para a modelagem estatística foram usados o algoritmo Naive Bayes, o Modelo Linear Geral, a Regressão Logística, o Fast Large Margin e a Árvore de Decisão com o apoio da ferramenta RapidMiner. Entre os principais resultados encontra-se o facto de as páginas dos cursos e de suas unidades curriculares atraírem um maior volume dos usuários, mas em que eles têm um menor engajamento. Por outro lado, as páginas com conteúdos mais generalistas atraem usuários que resultam em maior engajamento, assim como as páginas dos cursos de mestrado e de outros cursos da instituição. Este trabalho deixa abertura para a investigação de como aumentar o engajamento dos usuários que chegam pelas páginas que atraem mais volume de usuários e sobre como o reconhecimento da marca pode impulsionar a atração dos usuários pelas páginas que geralmente levam a maiores níveis de engajamento.

Palavras-chave: Educação Superior, Marketing Digital, Web Analytics, Data Mining, Machine Learning, Experiência do cliente, Web Design, CRISP-DM



Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

# ABSTRACT

When choosing a higher education course to enrol, future students interact with several institutional channels. Among them there is the website, which can incentivize or discourage prospective students by how engaged they get with its content. Visitor's behaviour is monitored by collecting data in their interactions using Web Analytics Tools, which allow their analysis and, consequently, generating insights to improve visitor's engagement. This study proposes the use of Data Mining techniques for insight generations focused on the undergraduate courses of a Portuguese higher education institution supported by the CRISP-DM methodology. The collection, storage and transformation have been performed using Google Tag Manager, Analytics, BigQuery and Data Studio. Next, the data have been statistically modelled using the algorithms Naive Bayes, Linear General Model, Logistic Regression, Fast Large Margin and Decision Tree with the support of RapidMiner. The main results of this results pointed that: course pages and course unit pages attract a larger volume of users, but they show lower engagement levels; more generalist content pages, as the website's homepage and the overall undergraduate course introduction page as much as masters' course pages and other course pages. This work leaves questions that can be further investigated as how to increase of user engagement of course pages and course unit pages and on the power of brand awareness when attracting users to more general pages that generate higher engagement levels.

Keywords: Higher Education, Digital Marketing, Web Analytics, Data Mining, Machine Learning, Customer Experience, Web Design, CRISP-DM

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

# INDEX OF FIGURES





Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

# INDEX OF TABLES




Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

# LIST OF ABBREVIATIONS AND ACRONYMS

HEI – Higher Education Institution

IES – Instituição de Ensino Superior

CRM – Customer Relashionship Management

GCII - Gabinete de Comunicação Institucional e Imagem

DCTI - Departamento de Tecnologias de Informação e Comunicação

ROC - Receiver Operator Characteristic Curve

AUC – Area Under the ROC curve


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

# INTRODUCTION

Workers, as part of the labour market, try to stand out by enrolling courses from Higher Education Institutions (HEIs) and demanding upper quality from them. Consequently, HEIs are incentivized to compete to offer what they look for (Nuriadi, 2021).

Increasing the number of course applicants is a key objective that impacts directly on higher revenue levels, notoriety and quality of student body in a HEI.

Some of the strategies to draw attention to the institution can be building and communicating about university facilities, teaching quality, affordable institutions fees among other added-values techniques. Gaining public trust might require improvements on the positive perception of the HEIs brand image, in special parents who are planning to send their children to college. Adapting and updating its positioning in each channel constantly ask for a previous understanding of how a product or service is accepted or used by the public (Nuriadi, 2021).

Websites appear as a relevant channel considering that this technology have impacted human activities and interactions on a large-scale according to Tiago and Verissimo (2014). Therefore, it is an important interface that HEIs need to consider positioning their marketing activities digitally. To understand and to draw new marketing strategies Business Intelligence and Customer Insights help on the investigation of visitor's interactions using massive data created in this environment. In this context, data mining can be helpful to perform efficient and insightful analysis (Osman, 2019).

A HEI with typical opportunities and challenges of the sector on Digital Marketing has been chosen as a case-study, focused on undergraduate programs. It is sited in the centre of Portugal and offers courses in diverse fields of study such as agriculture, health, business, engineering, education and arts from technical to master's programs.

The project general goal was to increase enrolment intentions — specifically, the proportion of users who checked course pages — by incrementing the number of applications for the institution's courses. With this purpose, understanding the users' journey from their first contact with the HEI's brand, the challenges and values perception they find might be beneficial in achieving this goal (Kalbach, 2021). This objective has been explored by Pinto et al. (2023a, 2023b) and expanded in this study.


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

To set precise objectives, the digital environment of the institution were both thoroughly examined. The case-study HEI's internal data was mapped, scoped, gathered, stored, and examined using data mining tools and visualizations. Then, groups were compared and patterns of visitors who accomplished the intended goal and those who did not. The project was carried out using the CRISP-DM data mining methodology, which has proven to be adaptable enough to be applied in other sectors. Google Tag Manager was used to gather the data, Google Analytics and Google BigQuery were used to store and convert the data, and RapidMiner was utilized to perform statistical modelling.

In the “Literature Review” chapter, the essential concepts for the comprehension of this work are presented. It is initiated by the topics on Digital Marketing for HEIs, which clarify the relevance and applications of digital channels to enhance results of this type of organization. In sequence, the use of Customer Experience and Web Analytics to organize your digital efforts in each phase of the relationship between the brand and the user is treated in detail. Then, the concepts of Data Mining and Data Quality for understanding, preparing and analyzing the data are presented.

The “Methodology” chapter has the function of describing CRISP-DM phases and defining the details for the data preparation. It recalls concepts from the previous section brought to the case of this study. The following chapters cover each phase of the methodology.

In “Business understanding” the digital environment of the institutions was mapped and the behavior of the Portuguese user in HEI’s website is analyzed. It compasses listing and understanding the functionality of that tool, bringing third-party studies together with data collected from the main HEI players there.

In “Data understanding” the dimensions, attributes and metrics are described and assessed in terms data quality. In other words, each aspect of the raw data set is presented to identify the most representative numbers and the need to aggregate less relevant numbers.


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

"Data preparation" brings the data transformation needs addressed in the "Data understanding" receive solutions with the addition of new dimensions to categorize the dimensions and metrics that cannot be used for modelling.

Then, "Modeling" defines and justifies the use of each algorithm for the three main questions of this study: "Who does access the website?", "Who does see course pages?" and "Who does return to the website?".

In sequence, the "Evaluation" gathers the main results of the analysis. This means, presenting the characteristics of the users analyzed or from the groups identified. It is used to contrast with the information collected in the previous sections to reach clearer insights.

In the "Conclusion" the content of this study is revised and summarized pointing to the possibility of future studies.


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

# 1 LITERATURE REVIEW

HEIs can optimize their efforts for attracting and retaining students using digital channels, which became an essential part of our contemporary life. Digital Marketing strategies commonly applied in the sector are presented in the first part of this chapter to elucidate opportunities worth considering. To make it work successfully, it is important to orchestrate the digital assets with clear objectives under the light of your Customer Experience, fed specially by digital data as shown in sequence. In the last subdivision of this literature review, the process of Data Mining, namely contextualization, cleaning, preparation and modeling that data must go through to be as exploited as much as possible is introduced.

## 1.1 Digital Marketing for Higher Education Institutions

The widespread use of web technologies had a significant impact on human settings, leading to a demand to adjust in marketing strategies to consider digital channels (Tiago & Veríssimo, 2014).

Digital marketing can be described according to different approaches as a mean to contact consumers more effectively, with greater personalisation and at a lower cost, Nurtirtaway et al. (2021) concentrated on the innovativeness of moving beyond the conventional analogical interactions. According to Forghani et al. (2021), digital marketing originated with the internet and search engines but has now expanded to include email, websites, blogs, and social networking. Muller (2019) emphasizes the use of the internet for product promotion and sales digitally.

Targeting prospective students, enlisting student engagement, giving placements, crafting curricula, providing career counselling, developing alumni contacts, and building professional student networks are some areas where digital marketing can be useful in HEI (Harbi and Ali, 2022). Online marketing benefits academic communications by providing a collaborative two-way channel that transcends physical and temporal limitations (Bateman, 2021).

In a study, Del Rocio Bonilla et al (2020) showed that Instagram, an image and video-focused online social media, was more successful and efficient at attracting


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

students' attention than posters, events, and sports. Kusumawati (2019) highlighted the importance of using social media while deciding which degree to enrol in.

Rajkumar et al. (2021) emphasized the value of digital marketing in improving the outcomes of the admissions decision-making process for students. Digital media, it was noted, might involve "engaging user-generated content through text, images, infographics, videos, and podcasts" for potential applicants. In addition to analytical management, Oré Calixto (2021) emphasized the value of content marketing for the educational industry. Their investigation concentrated on how these tactics improved customer relationship management (CRM) processes. According to research by Basha (2019), 62% of students' decisions about higher education institutions were influenced by digital marketing.

The ability to gather additional information about user behaviour to develop optimization methods is another benefit of digital marketing. Business intelligence and customer insights are crucial for understanding and supporting new marketing engagements, according to Stone and Woodcock (2013). The key to understanding behaviours and defining improvement plans may lie in mapping their interactions and behaviours (Kalbach, 2021).

## 1.2 Customer Experience and Web Analytics

Understanding what can persuade customers to make a purchase is vital when managing a business with sales goals. The same is true for prospective students interested in enrolling in a course. In this project, consumers and students are compared as broader concepts. If they find the value they were looking for in an offer, they might decide to buy.

Sheth et al. (1991) suggested that some values might be functional, such as creating a quick, facilitated, and reliable process to get a higher education degree; social, such as interacting with people who share a lifestyle; emotional, such as satisfying family expectations of earning a higher education or feeling prepared for the competitive job market; epistemic, such as satisfying one's curiosity in a field of interest; and conditional, such as satisfying your expectations or needs regarding cost of


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

living, entry requirements, the specific topic they aim to pursue a career or other conditions that are specific for some prospective students.

It can be challenging to see and prioritize value-driving improvements across all an organization's channels, products, processes, departments, roles, measurements, strategies, and goals.

The elements of the business and the customer's journey can be matched by creating a customer journey. For instance, it can help in comprehending the effects of displaying important information for potential students only on less accessible website pages or of not advertising a specific event. It can assist with visualizing when and how much they provide in value perception to applicants at each phase of the application process for a course and the overall picture.

According to Kalbach (2021), there are four crucial components to define in a customer map. The point of view, which can be in this case the prospective student, the internal organizational department or platform which play as active player experiencing or providing the journey. The scope is the order, which includes the beginning and end of focus of the mapping. Another component is the focus, which defines the elements you want to evaluate in each of these processes as clients of various categories; channels like online ads, websites, social media, online or offline events; and internal departments like the university or school admissions office or financial teams. Finally, the structure of the diagram, which is the visual presentation framework, which can be a mental map, a table or a linear chronology among other formats.

According to Pedowitz Group's loop-shaped suggested customer experience, each stage of a prospective customer's relationship and experience can often go through several phases (Holiday, 2018).

The Loops is split into two more substantial, related sections: Client acquisition and retention. It is possible to safeguard and enhance value for the brand by studying the Customer Retention phase, which includes such phases as onboarding, adoption, value realization, loyalty, and advocacy. Whereas during the customer acquisition, which is the focus of this work, the customer can be in one of these phases:


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

- Unaware and Aware: the consumer interacted with the brand in some way, usually by seeing on numerous online and offline ads, reading comments from friends, or taking part in events. They are now familiar with the business, its personality, value proposition, and services and goods it offers. There isn't much interest in buying yet, though.
- Consideration and Evaluation: the prospective consumer has a problem to address, and they are aware of the kinds of goods and services they must purchase to support them thanks to their prior interactions with the brand. The options provided by your company, your rivals, and substitutes are now compared using various criteria that reflect their opinion of value, such as price, quality, velocity, and so forth.
- Decision: the customer has made the decision to buy the product at this point. Now, it's critical to remove the most typical obstacles that could cause customers to back out of the transaction.

An example of Kalbach's suggestions to create a table-shaped Customer Journey aligning the institution's platforms and internal departments can be seen in Figure 1. In this simple representation, the point of view of the general candidate can be seen together as the internal team interactions and their supporting platforms. Other aspects displayed is the scope, which in this case follows the Customer Acquisition phases according to the Loop (Holiday, 2018) as much as the focus on some values the candidate might have and their levels of satisfaction in each of them.

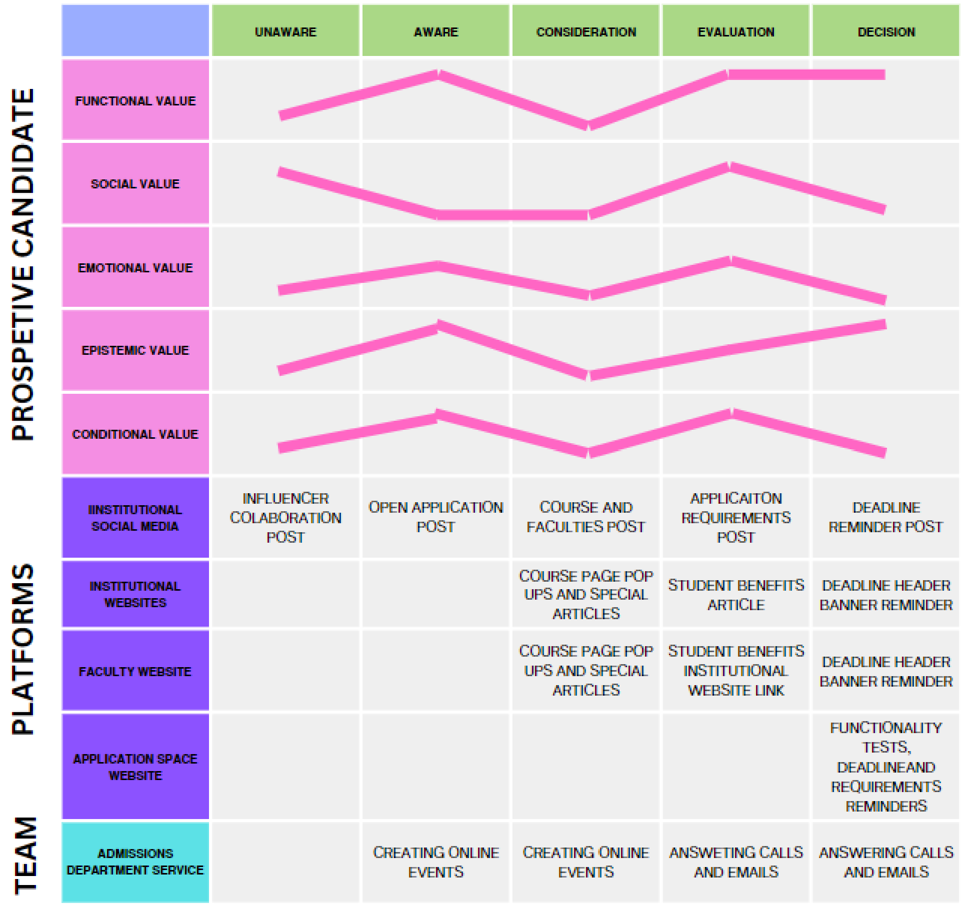


Figure 1 Sample Customer Journey Visualization

This Customer Journey is created using data from several sources. Data from users' behaviour both within and outside the website is needed to accurately describe the most relevant users and their behaviour throughout the entire journey.

Web analytics tools are used to analyse website data, and they are characterized as follows by Palomino et al. (2021):

> “Web analytics tools collect, analyse, and present web usage data, in ways that can highlight what on the website works well, and possibly can be leveraged further, and what might not work so well, and may require fixing, improvement, or removal. Raw usage data, such as web server access log files and click-stream logs, may contain too much information to be useful unprocessed for most purposes. Web analytics tools use metrics to summarize web usage data as simple numbers that are easy to work with.”

Web Analytics can be userful for generating audience segmentations. Donavan et al. (1999) describes that it is composed by three parts: finding homogeneus groups within heterogeneous groups, then choosing one or several of them and finally developing a program to serve the communication needs of that audience. As a result of this study it is possible to verify audiences that are generated by the applied models.

Data analysis can be a difficult undertaking given the variety of approaches available, particularly data mining techniques, which are discussed in the following section.

## 1.3 Data Mining Techniques and Data Quality

The amount of data gathered and stored increases quickly as technology advance. Extracting useful insights from them is now thought to be a crucial component of any brand's growth. (Osman, 2019). The use of collecting, extraction, analysis, and statistical methodologies along with data mining techniques is suggested to successfully comprehend and study these massive data sources. To choose the appropriate strategy to use, it is crucial to map the desired outcome.

The goal of the association approach is to identify the connection between variables contained within the same datasets. It would be possible to see, for instance, which courses users from each region see more often.

The classification method to segregate groups in the dataset to reach a goal. Its objective is to forecast how the ensuing cases will behave. It might be used, for instance, to identify individuals who might visit the website again to view other courses.

By identifying the most frequent behaviours, the clustering technique aims to categorize dataset examples into groups based on their similarity.

Decision trees examine which combinations of attributes have the most influence on a chosen target and prioritize the most pertinent attributes, such as the set of factors that contribute to a session with higher pageviews.

How many courses a user can see is an example of a prediction, which is a strategy that examines other variables to propose which value might occur given a combination of the others.

Neural networks are a technique that can artificially produce a more complicated set of layers of rules to suggest an attribute based on an example-case.

These methods can be used isolated or in combination to produce insightful results. The final quality of the analysis can be improved by selecting them and putting them into practice while employing an orientation strategy.

These models work with the use of prepared data to meet adequate quality levels. In this case, quality can be measured with variables as ID-ness, Missing Values, Stability, and Correlation were used to evaluate the data quality. According to Mandalapu and Gong (2019), these are:

> "Correlation in this study considers the linear correlation between attribute and target column. The percentage of ID-ness implies in the percentage of different values present in a column. For instance, an attribute with incremental values can have an ID-ness of 100%. Stability is the measure of constant values in an attribute. Stability is zero if there are no similar values where stability is 100 percent of all the values are the same in an attribute. Missing value measure is the percentage of values missing in an attribute."


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

# 2 METHODOLOGY

Designing solutions for processing massive datasets have become more popular in data science. Gaining "valuable insights by mathematical and analytical models and applications" is the goal of this subject. These projects use process approaches to their advantage by focusing on success criteria. Since its debut in 2000, CRISP-DM has become a popular project methodology in large part because it is a de-facto standard, which implies that it is generally recognized, simple to use, well-organized, trustworthy, and a process model that is independent of any industry (Schroer et al., 2021).

CRISP-DM phases may have a non-linear nature, which means that more than one phase might be elaborated simultaneously as seen on Figure 2.

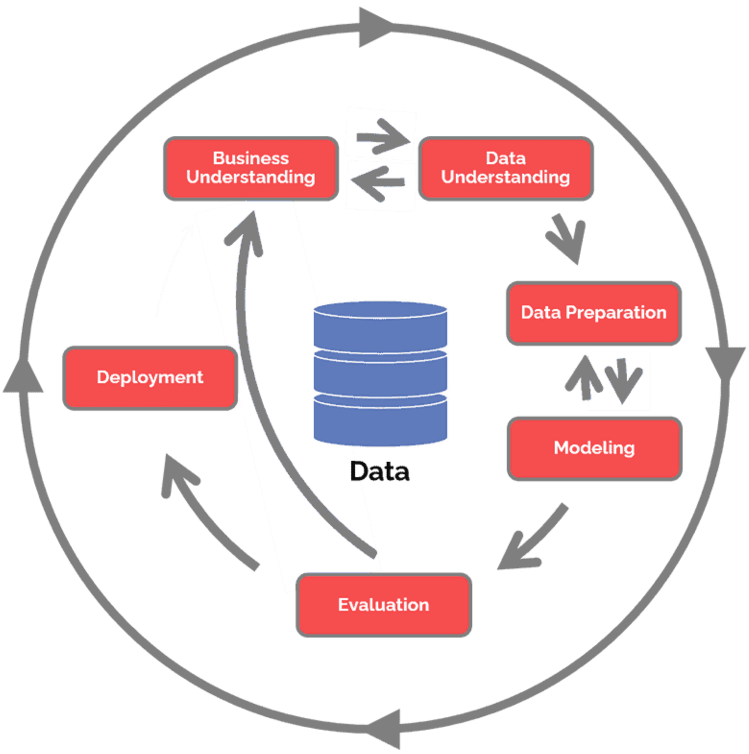


Figure 2 CRISP-DM Phases

Source: Hotz (2023)

The CRISP-DM project is divided into six parts. The first stage, business understanding, involves evaluating the current state of the company, outlining the data sources that are accessible, discussing specific data mining objectives and methods, and creating a project schedule.

To meet the need for data transformations prior to applying the models, the data understanding phase refers to understanding the data source, extracting the data, describing it, and evaluating its quality. A sufficient set of attributes for analysis were those with ID-ness and stability lower than 90%, correlation between 1% and 40%, and fewer than 70% of missing values.

Then, during the data preparation stage, the dataset is cleaned up by removing low quality features, adding new attributes and categories, and integrating tables following steps of a pipeline.

The modelling technique is chosen in the modelling phase with sufficient reasons taking evaluation criteria and established parameter settings into account. It is possible to use multiple approaches.

The X-Means approach was used for clustering to determine which user profiles are the most prevalent. It is an evolution of the K-means clustering in which the number of clusters was automatically recognized up to a maximum previously established number, as opposed to the first, which adhered to a defined initial number of clusters. It begins by selecting the fewest clusters possible within a specified range and then calculates a Bayesian Information Criterion (BIC) score, which assesses the likelihood of each cluster to share similarities. This likelihood quantifies how closely each example is categorized such that it is possible that they are all members of the same category. Until the maximum number, those with poorer outcomes are separated into new clusters (Yassir et al., 2020).

Nave Bayes, Generalized Linear Model, Logistic Regression, Fast Large Margin, and Decision Tree were the algorithms and methods used for classification. In contrast to the Generalized Linear Model, which is used to determine the correlation between two or more variables with cause-and-effect relations, Nave Bayes is "a simple probabilistic classifier that calculate a set of probabilities by summing the frequency and value combination of given dataset" (Peling et al., 2017). Based on a Support Vector Machine, the Fast Large Margin can handle a wide range of attributes (Saputro et al.,



Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

2021). Decision trees choose the set of characteristics that have the greatest impact on the chosen targeted variable (Osman, 2019).

The Area Under the Curve of Receiver Operating Characteristics (ROC's AUC) was used to assess which model to use because it aggregates various analyses from the confusion matrix, including sensitivity, which measures the proportion of correctly classified positives, and specificity, which assesses the proportion of correctly classified negatives (Zhu et al., 2010).

A confusion matrix shows the number of cases that have been correctly or incorrectly classified as positive or negative after applying a prediction model on a sample as displayed on Figure 3.

The sensibility is an indicator created from the confusion matrix that aims to evaluation the proportion of positive cases that have been correctly classified by dividing them from all verified positive cases. It is considered that the higher the sensibility, the lower the probability of finding cases that have been predicted as positive but that were in fact negative.

The specificity is another measure based on the confusion matrix, which aims to assess if the model might classify incorrectly negative cases. It is calculated by dividing the number of cases that were negative and that were correctly classified as negative by the total number of predicted negative cases. This ratio should indicate that there is a lower possibility of predicting a case as negative when it is positive.

| | | **Actual** | |
|---|---|---|---|
| | | **Positive** | **Negative** |
| **Predicted** | **Positive** | True Positives (TP) | False Positives (FP) |
| | **Negative** | False Negatives (FN) | True Negatives (TN) |

Figure 3: Confusion matrix

The ROC's AUC aims to indicate the classification models that can more effectively identify correctly both positive and negative cases considering sensibility


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

and specificity equally. To do so, it considers the results of several tests to indicate those that performed better. A simpler way to understand it can be by seen it in a graph as shown in Figure 4. The results of the sensibility and of 1 minus specificity would represent the line in green and area under this graph would be the AUC. This means that the more the line is curved above the average line, the higher the sensibility and the specificity.

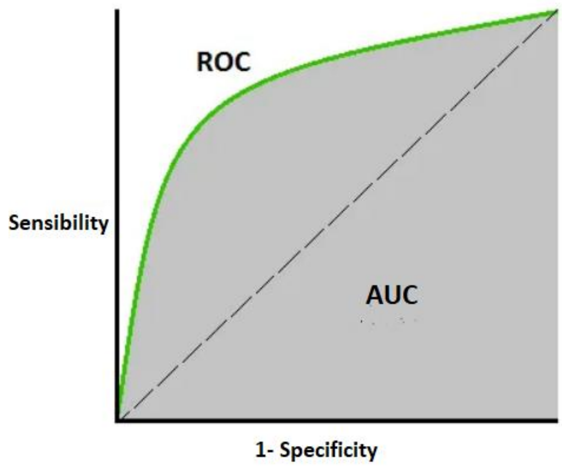


Figure 4: ROC's AUC

After choosing the most appropriate models, the outcomes are discussed while contrasting business objectives in the evaluation step.

The deployment of a user manual, software components, and the scheduling of performance monitoring are all possible activities that can be part of the deployment step. The deployment of this model will be an option for the institution in the future with further consideration of proper and efficient implementation and maintenance.


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

# 3 BUSINESS UNDERSTANDING

The top 10 Portuguese HEI presented around 18 million sessions in November 2022. The first university counts for 23% of them in number of sessions in November 2022 according to SimilarWeb, a platform about competitor's website information. The case-study university is ranked in 10th place in number of sessions in the period accounting for 4% of the sessions, similar to the 8th and the 9th place, with 5% each of them as seen in Figure 5 (SimilarWeb, 2022).

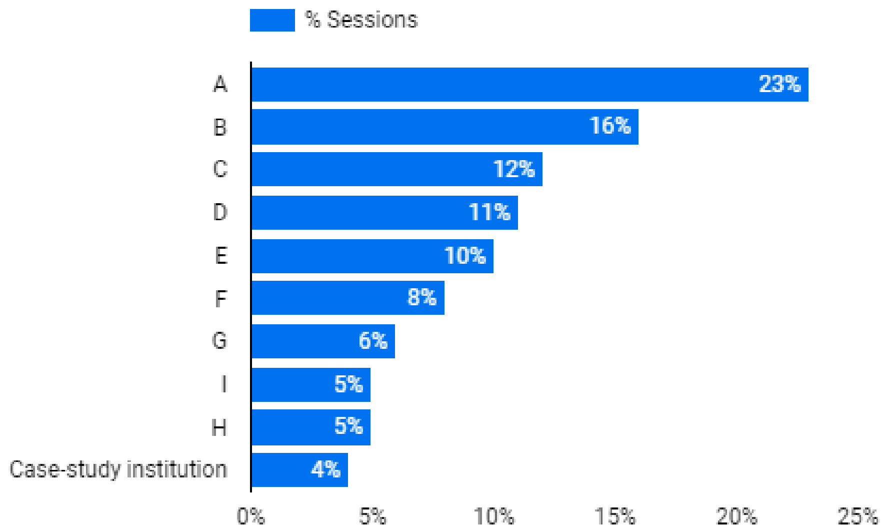


Figure 5: Percentage of Sessions from the top 10 HEI's websites in Portugal

Source: SimilarWeb (2022)

In terms of session length, the 10 Portuguese HEI sessions spent 2,3 times longer than the global average of 2:25 minutes (Bailyn, 2022), which may indicate that Portuguese HEI audiences value online content more than audiences in other geographic areas.

The case-study HEI achieved the 7th place out of the 10 top Portuguese HEI websites in terms of average session duration and average number of pageviews per


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

session as shown in Figure 6. It demonstrates that it is possible to generate more engaged experiences on the website. As seen that the Portuguese users apparently value the website content spending more time than the average in the world, it might indicate that improving the content of the case-study HEI can lead to positive results in the value of the brand.

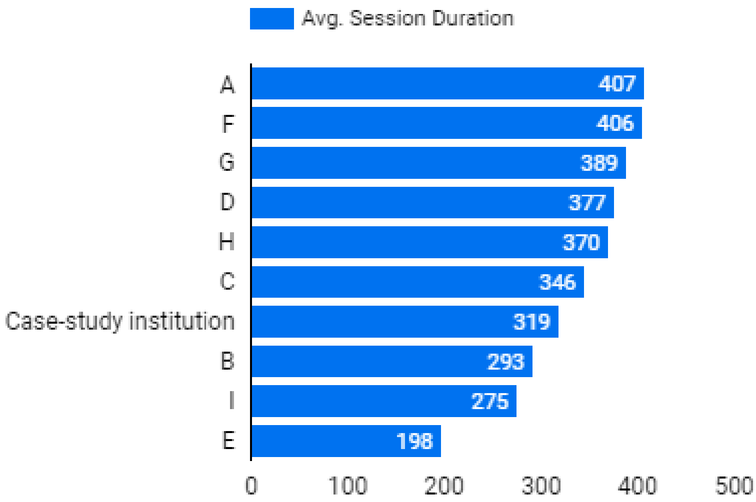


Figure 6: Average Session Duration by HEI Website in seconds

Source: Similarweb (2022)

In addition to the main website, the case-study institution has several other online and offline channels for communicating with prospective and current students, including the academic service offices of each faculty (both in-person and via email), the university communications office email, higher education events, the university website, the student space website, websites for each of the six websites, and their respective social media accounts on Facebook and Instagram.

Beyond the central website, the case-study institution counts with several online and offline channels of contact with prospective and current students such as the academic service department of each faculty onsite and by email, the university communications department email, higher education events, the university website, the student space website, websites of each of the six websites and their respective social media profiles on Facebook and Instagram.

The website offers information on the university, its programs, internal groups, activities, locale, amenities, and services. The email address and phone number for individuals who are interested are presented for more information. When the study was conducted, it was not possible to directly gather the contact information of visitors who were most interested in the course as a following step to provide a closer approach to them. It did not allow visitors who were interested in a course to actively leave a note or ask to be contacted by a school official to learn more about it. Then, from the view of a prospective student, it can be categorized as a channel used during the consideration and stage. Its goal is to increase user involvement, much like a showroom.

It was decided to simply use the central website as a data source to be evaluated, keeping in mind that it is the most significant in terms of user volume, as the institution is beginning to use data to define their digital strategy with this project.

This study concentrated on undergraduate courses, which had 31% of the most active users among all users who saw a course's pages throughout the study period.

For this study, the period of April to June 2023 was chosen since it can be seen as long enough, yet brief enough to produce the initial insights with the potential to be expanded in subsequent studies.

People who have seen pages about undergraduate programs are considered the most interested in them in this study. The objectives of this project included: using cluster analysis, identifying the most frequent users and their online behaviours; using classification analysis, comparing the characteristics of sessions of those who check course content more frequently with those who didn't; and using classification analysis, comparing sessions of those who check course-related content and return for more sessions.

The course page views that outline the course objectives, enrolment requirements, and curriculum modules were designated as the goal-content. The increase in the percentage of users who saw course pages during the period relative to all users is the goal of this study. It was perceived that 31% of the users have seen the undergraduate page as seen of Figure 7.


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

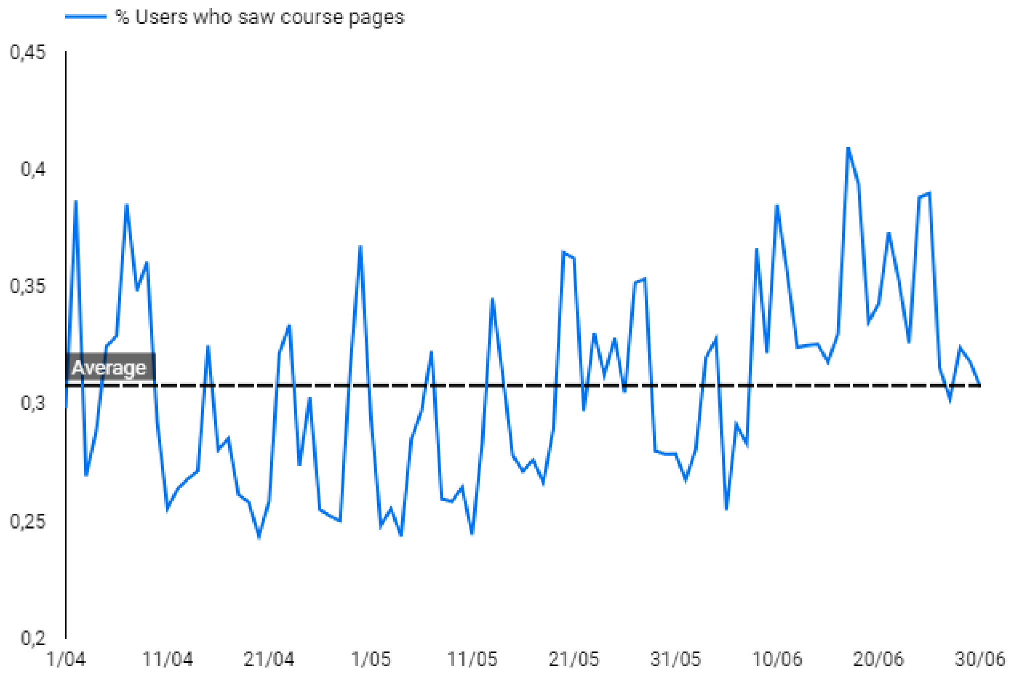


Figure 7: Evolution of the Percentage of users who saw course pages

Source: Own elaboration using the Google Analytics institutional tool

The goal of the Data Understanding phase was to extract reports from the tool and use Google BigQuery to map and visualize the dimensions and metrics in Google Analytics with more details. By examining visualizations and correlations, the limitations of data extraction, relevance, and the quality of those metrics and dimensions would be evaluated in turn to address the need for eliminating or transforming them.

To address the need for data transformations before applying algorithms, data understanding entails understanding the data source, extracting the data, describing it, and evaluating its quality. The indicators ID-ness, Missing Values, Stability, and Correlation were used to assess the quality of the data.

The impact of the extraction limitations on the need of transforming the source data to produce the final dataset suitable for modelling was addressed at this stage. The pipeline for the data transformation strategy was established during the data preparation to provide a final dataset that is integrated and can be algorithmically modelled. The algorithms are then applied to the final dataset during the modelling step, and the outcomes are then made available. In the evaluation, the models' error levels are compared to determine which one has achieved the highest levels of correctness.

# 4 DATA UNDERSTANDING

In the "Data Understanding" phase it is identified the volume of data, the dimensions and metrics accordingly with the assessment of data quality.

The Google Analytics account of the case-institution is the web analytics tool used as data source. It counts with storage data regarding user variable types as user acquisition, device, content, geography, demographics, sessions, users, events and among others which are measured as number of users, of sessions and events, including pageviews. The total number of events was 1855874, and there were 140781 engaged sessions, which means that are those who spent 10 seconds or more in a session, who performed 1 conversion event or that saw 2 or more pages or in their first visit (Analytics Help, 2023a, 2023b), and 123724 active users, who have engaged sessions or thar had their first visit in the website (Analytics Help, 2023), in the period. Eighty two percent of the people who visited the website came from Portugal. As a first phase, these were selected as the only ones filtered out of the analysis. The study of prospective international students can be a possible future study. These data can be seen on Figure 8.

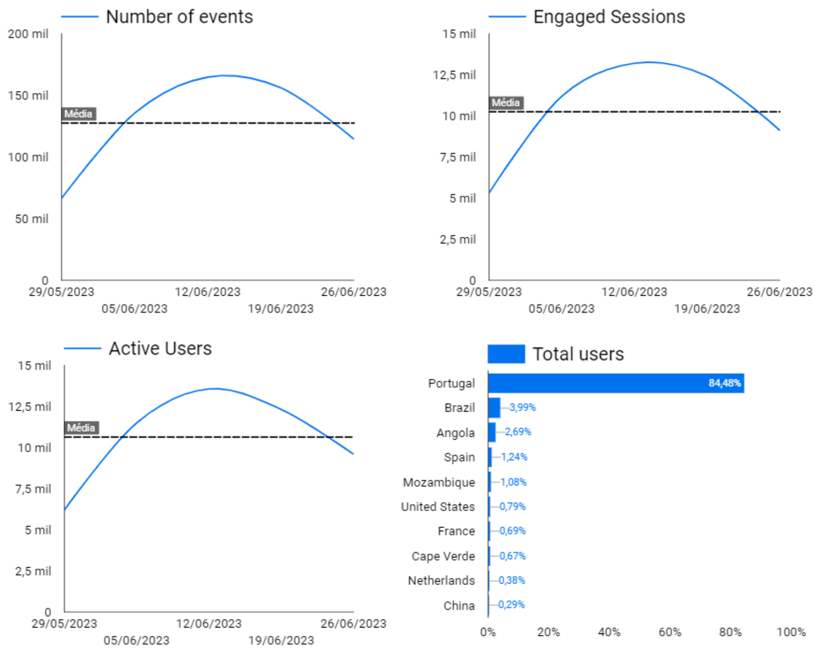


Figure 8: Key volume metrics of the dataset

Source: Own elaboration using the Google Analytics institutional tool


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

The weekly evolution of users revealed a number of users typically between 10,000 and 15,000 every period, stable engaged sessions per user of about 1, and pageviews per user of approximately 4 and more.

For purposes of comparison, these indicators were combined into a group of users concentrating on:

- Device category, specifically, desktop, mobile, tablet, and smart tv, with 50%, 49%, 9%, and 1% of users each as seen of Figure 9.

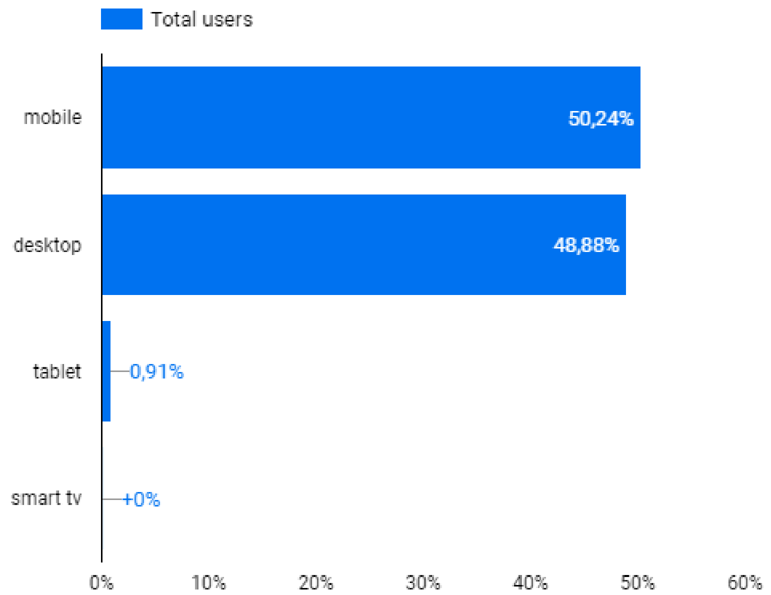


Figure 9: Users by Device Category

Source: Own elaboration using the Google Analytics institutional tool

- Region, Portugal users, who make up 82,3% of users in the area where the user completed the session that was restricted to them, are the target audience for the courses. Accordingly, "Lisbon" (35%), "Coimbra District" (20%), "Porto District" (16%), "Setubal" (13%), "Aveiro District" (5%), "Leiria District" (4%), "Braga" (3,5%) and others (3%), are the regions with the highest concentration of active users as seen on Figure 10.


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

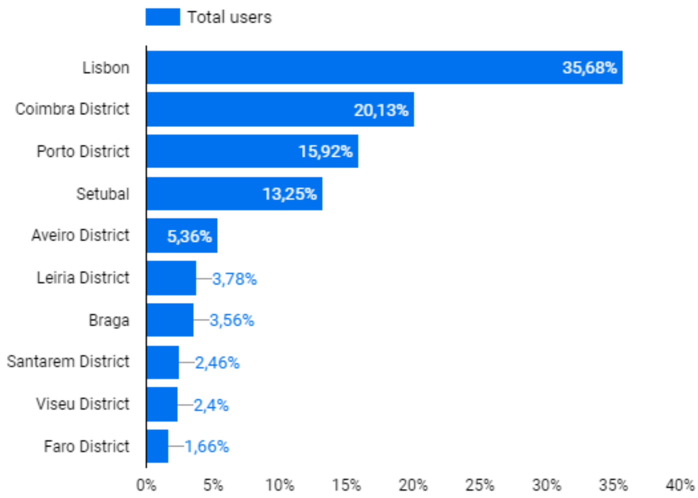


Figure 10: Users by Region

Source: Own elaboration using the Google Analytics institutional tool

- Medium of the first session, is if the URL was typed directly into the browser search bar as stated by (none) when entering the website for the first time or when using a link on an external website to reach the website for the first time. as the organic searching engine used to reach the page. In terms of the total number of active users, "organic" (78%), "(none)" (11%), "referral" (11%), "email" (less than 1%), and "Website" (less than 1%) are the most representative mediums of the first session as seen on Figure 11.

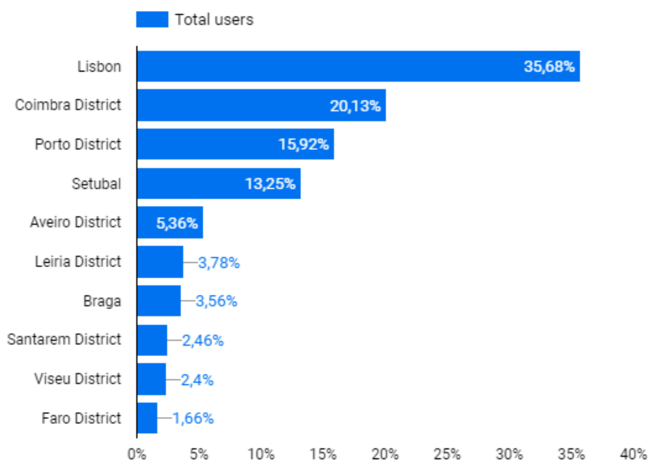


Figure 11: Users by Medium of the first session

Source: Own elaboration using the Google Analytics institutional tool

- Source of the first session, when entering the website for the first time, as the previous website, the searching engine used to achieve the website, or if the URL has been put directly on the browser search bar as specified by (direct). The terms "google" (74%), "(direct)" (11%), "esec.pt" (6%), "bing" (4%), "isec.pt" (2%) and others (3%), are among the top values for this dimension as seen on Figure 12.

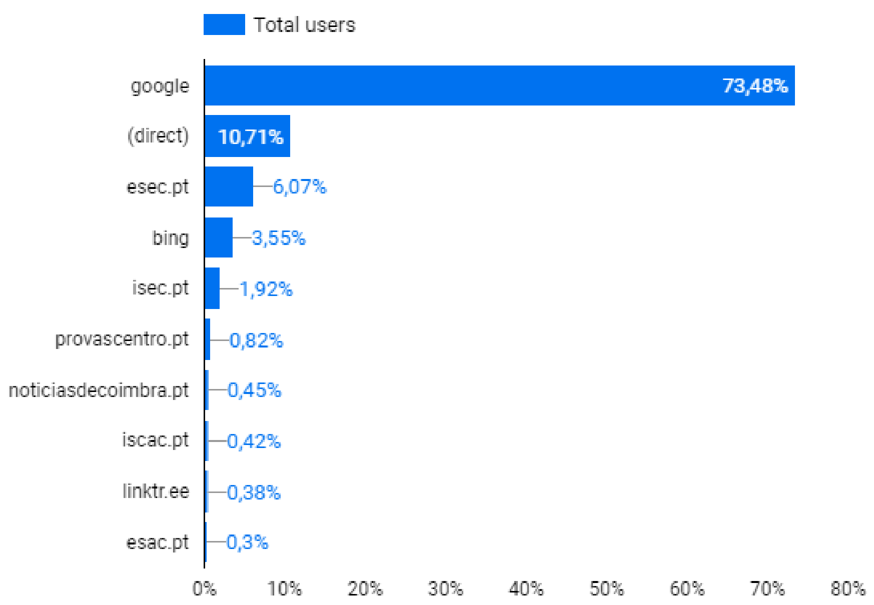


Figure 12: Users by source of the first session

Source: Own elaboration using the Google Analytics institutional tool

- Medium of the session, it takes into account the medium that was employed during the session in contrast to the Medium of the first session. Accordingly, the groups "organic" (75%), "referral" (13%), "(none)" (11,85%), and others (0,15%) have the highest percentages of active users as seen on Figure 13.

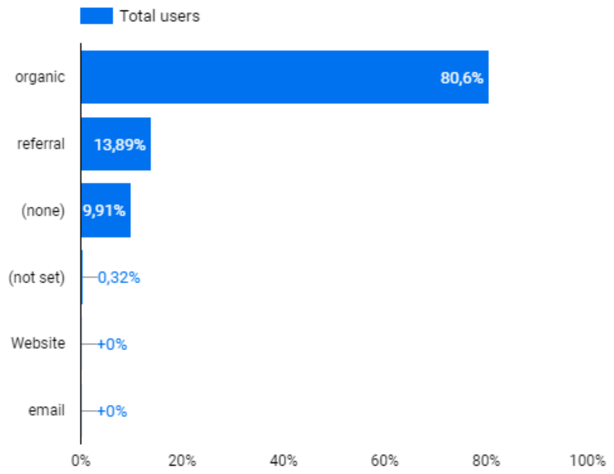


Figure 13: Users by medium of the first session

Source: Own elaboration using the Google Analytics institutional tool

- Source of the session, it considers the source used in the session itself in comparison to the source of the first session. The terms "google" (70%), "(direct)" (12%), "education school website" (6%) and others (12%) were the most frequently used values as seen on Figure 14.

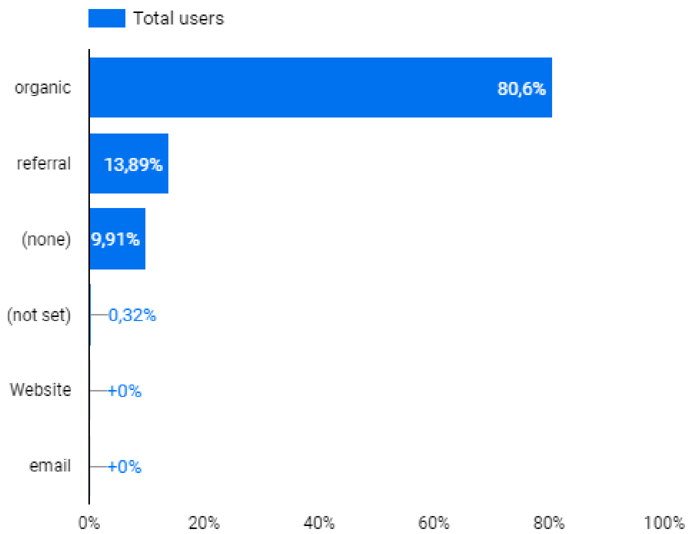


Figure 14: Users by Source of the session

Source: Own elaboration using the Google Analytics institutional tool


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

- Landing pages, via the link they clicked on or the URL they searched in the browser, as the first page they reach on the website. The "Homepage" (10,3%), the contact page for the health school (3,3%), the general page for technical courses (3,05%), the page for the university canteens (2,9%), the page with the complete list of courses offered by the university (2,48%), and other less popular landing pages (77,7%) are the landing page URLs with the most users as seen on Figure 15.

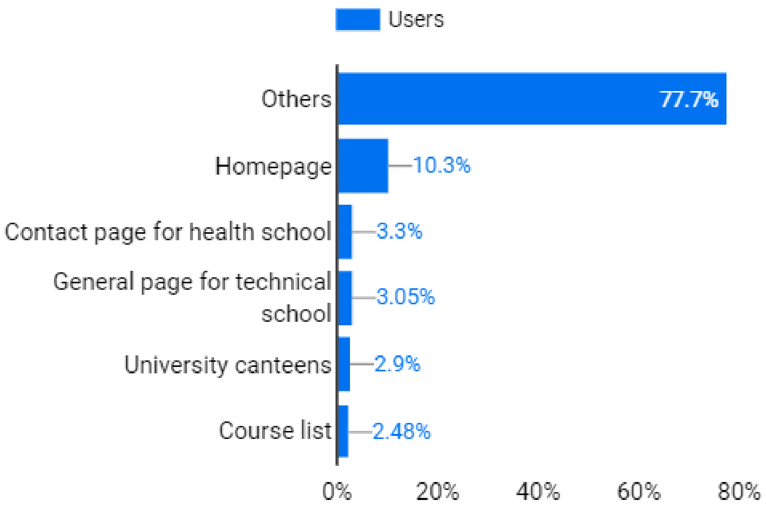


Figure 15: Users by Landing page category

Source: Own elaboration using the Google Analytics institutional tool

The number of values, percentage of missing values, ID-ness, and stability have all been used to analyse the quality of each dimension. More than 150 values were provided in the landing page, the source of the initial session, and the source of the session, which may indicate the need to categorize values. Higher degrees of stability were displayed by the "Medium of the first the session" and "Medium of the session," suggesting that less important values might be categorized as "others" in these cases. No dimension has missing values totalling more than 0,56% from its source. Given that they make up 21 regions, grouping regions is an option. The analysis is shown in Table 1.

The values with fewer active users were grouped together to reduce the number of values for "Region", "Source of the first session," and "Source of the session," and a categorization was made for "Landing Page's Page Path" based on an analysis of the website and the URLs, as further explained in the section on data preparation.

The metrics have undergone the same study of the variables' quality. As shown in Table 2, where the departure from the average number of users per group is 5,2, it is feasible to see that some groups have a concentration of users.

The main positive correlations for the "Course Offer Page" and "No Course" filters were 0,73 and 0,82, respectively, when the search feature was used, according to the analysis of correlations over 0,4. However, the filter that is most frequently used is to only display "technical courses" (correlation = 0,63).

| Dimensions | Number of values | ID-ness | Stability |
|---|---|---|---|
| Region | 21 | < 1% | 19% |
| Source of the first session | 154 | < 1% | 58% |
| Medium of the first session | 5 | < 1% | 65% |
| Source of the session | 173 | < 1% | 63% |
| Medium of the session | 5 | < 1% | 72% |
| Landing Page's Page Path | 5341 | 14% | 2% |
| Device Category | 3 | < 1% | 60% |

Table 1: Quality of data of dimensions

Source: Own elaboration using the Google Analytics institutional tool

Undergraduate courses are the most frequently accessed "Type of course" page (correlation = 0,75). In comparison to other types of courses, the undergraduate courses appear to be the principal course pages utilized as landing pages (correlation = 0,58). The groups with the most active users are often the ones that view the most pages (correlation of 0,72), the most sessions (correlation of 0,54), and the least amount of course pages (correlation of 0,6). This displays a profile of users who engage in more sessions and view more pages but don't view course pages. The majority of first-time users prefer to do sessions with the same medium (correlations, respectively, 0,65, 0,61, 0,52) whether they arrive directly, organically, or through a recommendation.

| Metrics | Average | Maximum | Deviation | Stability | Deviation/ Average |
|---|---|---|---|---|---|
| Active users - General | 3,728 | 1298 | 19,5 | 62% | 5,2 |
| Avg. Sessions per user - General | 0,789 | 45 | 0,7 | 55,00% | 0,9 |
| Avg. Pageviews per user - General | 3,179 | 390 | 6 | 43% | 1,9 |
| Percentage of users who checked course pages | 0,31 | 0,31 | 0,4 | 62% | 1,4 |

Table 2: Quality of data of metrics

Source: Own elaboration using the Google Analytics institutional tool

The undergraduate courses offered by the engineering school are the most often utilized course pages as landing pages, followed by those offered by the education school (correlations = 0,5 and 0,4 respectively).

Also described as relevant were the negative correlations below -0,4. With a poor correlation of roughly -0,96 between the experiences on desktop and mobile, it was clear that there were significant differences between the two. The "Medium of the first session" and the "Medium of the session" also shown that the values "(none)", "organic", and "referral" behave differently with adverse correlations lower than -0,4.

It was not found any correlation that explained clearly what causes higher percentages of users who saw course pages, and thus, it is needed to make further analysis with multivariate methods to get to them. Positive correlations might indicate groups of attributed that could be merged and the negative correlations might show which should not be due to divergent behaviours.



Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

# 5 DATA PREPARATION

By removing less significant variables, aggregating less significant features, or performing samplings, the data preparation stage seeks to improve the quality of the variables used in terms of relevance determined in the data understanding stage.

A large number of values were displayed for the Landing Page - Page attribute. These values were grouped by spotting trends in the URLs and by comprehending the website's sections. The page paths of the URLs, which are the passages between "/" symbols, were then divided into columns. The only Page Paths with stability lower than 90% and fewer than 70% of missing values were 2, 3, and 4. The categories of the variable Landing Page have been used in place of it in the modelling dataset. It was feasible to find 4 tiers of categories, each counting 20, 30, 31 and 26 values, by analysing the website's pages.

With the exception of "Medium of the session" being "(none)" because it showed a positive correlation with the value "(none)" of the "Medium of the first visit" and a negative correlation with the attribute’s "referral" and "organic" of the same variable, the values with less active users were then grouped in "others" up to a maximum of 15%. Other notable exceptions included the variables "Page Subcategory 1" values "education school" and "engineering school," which displayed favourable correlations with the value "Course Page" for "Page Category."


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

# 6 MODELLING

By removing less significant factors, the data preparation stage intended to increase the quality of the variables used in terms of significance found in the data understanding stage. In sequence, the algorithms are applied on data that is ready to be modelled. First, X-Means have been used for as a clustering method followed by the algorithms of Naïve Bayes, the Generalized Linear Model, Logistic Regression, Fast Large Margin and Decision Tree.

## 6.1 Who are the users of the website?

To identify who are the user of the website it was applied a clustering using the K-means and X-Means, which automatically identifies the number of clusters.

## 6.2 Who does access course pages?

The technique used was the classification of users into those who viewed more than 50% of the course pages and those who viewed less.

The "percentage of users who viewed course pages in comparison to all users" was the objective. The group of sessions for every combination of attributes served as the analytical unit. To ensure that the study was not driven by the volume of traffic but rather by the characteristics of the group, the variable "total active users" was removed.

The Generalized Linear Model and Logistic Regression had the highest AUC (0,883). Both had a same specificity of 53% and identical sensitivity (99,3%), however the first one had a 0,1% higher value. As seen in table 3, this model had an accuracy of 86%, and it was ultimately chosen.



Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

| | AUC | Sensitivity | Specificity | Accuracy |
|---|---|---|---|---|
| Naive Bayes | 0,839 | 75,1% | 76,7% | 75,6% |
| Generalized Linear Model | 0,883 | 99,3% | 52,7% | 86,0% |
| Logistic Regression | 0,883 | 99,3% | 52,6% | 85,9% |
| Fast Large Margin | 0,867 | 97,4% | 55,7% | 85,5% |
| Decision Tree | 0,814 | 99,9% | 1,7% | 71,8% |

Table 3: Measures of error from each method

## 6.3 Who saw course pages and returned to more visits on the website?

The classification approach has been used, with the average number of sessions per user seeing course pages set as the target. Classes were established as two equally sized dataset portions partitioned as lower outcomes for the desired measure. The second class, which includes the users with the most sessions, was designated as the one of higher interest. When examining this objective, the dimensions and metrics with the lowest quality were disregarded. Due to having more than 70% of the data missing, they were excluded to as "Page Path 1", "Page Path 4", "Page Path 5", "Page Path 6", "Page Path 7", and "Page Path 8" in this instance. To reduce the impact of user volume on the analysis, the metric "Active users" has also been removed. Both the "Page Title" and the "Avg. number of sessions per user - General" have been eliminated since any correlation between the targets and then wouldn't produce insights. The "Page Title" has been excluded because its text-ness is more than 66,06%.


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

# 7 EVALUATION

This chapter aims to bring the results from the Models applied and to confront it against the findings and the objective set on “Business Understanding”. The results coming from the models chosen for each question in the last chapter are going to be evaluated in the following sections.

## 7.1 Who are the users of the website?

Four of the seven clusters that the X-Means algorithm discovered comprised more than 5% of the active users. They are the main subject of this analysis. The centroid table's findings, which include the average value assigned to each metric in the model and the weight of each dimension for the dimensions, have been used to compare them.

The significant characteristics that they shared included the fact that they frequently arrived utilizing "google" as an organic search engine, which accounts for 60% to 70% of all active users in the four most significant clusters. Given that 30% of all active users are located in Lisbon, it looks to be the most pertinent region. Coimbra ranks in second place with at least 20% of them.

The cluster with the lowest percentage of active users who viewed course pages (9,68%) is cluster A, which accounts for 50% of all active users. They arrive via course unit pages (22,70%), course pages in general (22,70%), including master's, technical, and other courses (22,70%), and residual landing pages with few visitors, according to the Centroid table. It can suggest that, although course unit pages function as landing pages, they require further attention and a redesign in order to persuade users to view course pages.

Thirty percent of all active users were in the cluster B. They displayed the highest number of users who accessed pages for undergraduate courses (94,07%). They went straight to the course page for undergraduates. (80%), in particular for the education school (37%), which was followed by the engineering (14%) and health (13%) institutions. In contrast to cluster A, 26% of them arrive via "referral," which is when users access the case-HEI via links on other websites, and 14% of visitors arrive via the education school's website.



Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

The cluster C revealed an anticipated 34% of people who might view course pages and represents 15% of the users. Their primary characteristic is that 33% of visitors enter via the homepage, 17% via the contact page, and 13% via course pages. The most significant difference is that majority of these users (61%), as opposed to the other groups, arrive by mobile. It is clear to see that there is a higher percentage of users from the Coimbra District (27%) and Lisbon (50%) in this category.

Approximately 5.5% of users who are active are part of cluster D. Arrival via the homepage or the page that introduces all undergraduate courses (28%) is their most significant characteristic. Users from the Coimbra District make up a larger percentage (35%).

Even though clusters A and B get a larger number of users, it is clear from examining engagement indicators that these experiences do not engage users. Clusters C and D showed an average of 0,9 and 1,21 sessions per user, in contrast to clusters A and B which showed an average of 0,8 sessions, indicating that many sessions fall short of the required number of interactions to be considered engaged. The discrepancy is even more obvious when looking at "pages per user," which is 2,7 for clusters A and B while it is 4,8 and 11,55 for clusters C and D. This comparison might help indicating that traits more commonly found in the clusters C and D might be a better path for a user who is interested in courses such as arriving by the homepage or the general pages for the type of courses such as undergraduate courses, technical courses and master courses.

However, it could also be a sign that the path leading from undergraduate course pages and course unit pages needs to encourage users to view more content in a similar way to how clusters C and D do. Additionally, these groupings revealed a higher concentration of users from the Coimbra and Lisbon districts.

These results may also suggest that users are more accustomed to approaching the case-HEI by the names of the courses than by the brand name that refers to the entire university. We can see that the clusters that use the homepage and general course introductory pages the most typically view more content and visit the website more frequently. Because they arrived by clicking on search engine results that were


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

displayed without the aid of advertisements on these platforms, they are more familiar with the institutions' names.

In order to make it even better and try to implement this method in other areas, it can be an interesting next step to conduct additional research to determine why it occurs in Lisbon and the Coimbra District. As shown from the fact that individuals who are aware of the brand do spend more time on the website, it can also be fascinating to look into how to increase the institution's brand awareness in an effort to draw in more students.

## 7.2 Who does access course pages?

The attributes with positive and negative correlations are presented in this section. In particular, those from the schools of health, engineering, and business, followed by agriculture, and education, in that order, had landing pages that were undergraduate course pages (coefficient = -4.243). Their respective coefficients were -0,583, -0,377, -0,353, -0,116, and -0,078, respectively. But it is evident that, when compared to the "general undergraduate course page" (coefficient = 3,658), the course pages in particular do not bring engaged sessions. It is critical to think about upgrading the website's design because course pages constitute the most significant sort of landing page, accounting for up to 36% of all active users.

Users who visit course pages typically do so after finding them through Google's organic search (coefficient: 0,634) and occasionally through direct search (coefficient: 0,468).

According to correlations of 0,441 and 0,402, we can observe that the course pages for master’s programs, courses other than undergraduate, and technical programs draw more users than undergraduate course pages. Though they don't directly address the main issue, it can be an indication that those courses might be used as an engaging communication topic to draw students to undergraduate courses. Because they fall below a threshold of sessions, results from residual groups of sessions were not examined.

Sessions that see course pages do not appear to be helped by the "referral" source (coefficient = -0,466), particularly those that were originated from the education school


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

website (coefficient = -0,37). But it's crucial to note that, in comparison to the other schools, this one attracts the most visitors to its website even though they are less engaged.

## 7.3 Who saw course pages and returned to more visits on the website?

Each attribute can contribute positively or negatively to be classified as the group with higher tendencies to do more than one session or as the group with less chances to return to the website within the analysed period. In other words, those attributes that mean a higher influence in comparison to its peer values were considered the most relevant for this analysis. The most relevant outcome was that Coimbra users might return (20), more than users who started by any page related to undergraduate course content, with the exception of "Course unit pages" (-20), users who arrive by links of the education school (12), or users who come by any other link from another website to, considered as a source as "referral" (10). The likelihood of bringing individuals who view course pages and return was also demonstrated by the "Source of the first visit" being "direct" (12). When there are more pageviews per session, people are more likely to return to the website (10). Users who appear to be more likely to return to the website again seem to land on the "studying" area, the "homepage," the "Type of course" section, and the course pages from the education school (9, 8, 6, 6). People who access the website through the "Current student pages" "Page Category" can visit it more frequently than people who access it through the "health school courses" "Page Category." Important pages include those for "International Students" as well as the home page for visitors who arrive via "Contact Pages" (3) or return visits (4). Users who view the page using a desktop computer may come back (4), use technical course pages (3), or use "Course Offer Pages" (2) to get there. Users who access the website via "master's course pages" may not return (-3); similarly, users who access the site via organic search on their initial visit may not return as well (-11); or users who access the site via residual pages (-6) or visit sources (-7) or regions (-12).

Last but not least, it is possible to infer that the user who viewed the course pages and are most likely to return uses a desktop, is from the district of Coimbra, and arrives via "direct" or "referral," particularly from the education school website and landing


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

pages like those in the section "Studying," with the majority of users arriving by the "General Undergraduate course page" or by the "Homepage."


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

# 8 CONCLUSION

HEIs need to attract a large number of students to improve the quality of the student body, to generate more revenue and to achieve more notability. Digital marketing seems to be a possible solution. Contextualizing this solution requires considering the applicant's familiarity with data from sources like web analytics tools. Using the CRISP-DM approach, data mining can aid in obtaining insights from this data.

The website of a Portuguese HEI was the focus of this investigation. After learning about the company's digital channels and using tools like Google Tag Manager, Analytics, BigQuery, and RapidMiner to gather, prepare, and analyse the data using machine learning approaches to gain insights.

Users who arrive using "google" as a source, visitors who travel from Lisbon at a rate of about 30%, and visitors who originate in Coimbra at least 20% of the time were the most prevalent characteristics of users and their behaviour.

The users were grouped in the four most common clusters. Cluster A made up for 50% and represents the users who arrive mostly by course pages and course unit pages, but that did not present higher engagement. Cluster B accounted for 30% of users, who see more course pages. They typically arrive in course pages from the education school and occasionally via links from the faculties of the case-HEI. In sequence, Cluster C, reaching 15% of the users, might arrive using the homepage or the contact page, they might use mobile and do the visit from Coimbra or Lisbon. Their engagement level might be higher than those of Cluster A. The last cluster, D, showed 5,5% of users who arrive mostly from the “general undergraduate courses page” of from the home page and also bring higher engagement users when compared to Cluster A.

It was possible to see that the course page, which is the most typical landing page, can be improved to produce sessions that generate more engaged sessions. It was also noticed that the general undergraduate course page might draw students who are more interested in viewing course content, making it a good candidate for incentive as a landing page in campaigns on par with master courses and other courses. The users that return from organic search and directly type the website's URL into their browser are more likely to view course pages.

Some of the main indicators that a user who saw course pages and that might return to make more visits to the website were respectively: performing visits from Coimbra; those arriving by links in other websites, especially those who arrived by links from the education school website; those who views more pages; those who arrive on the homepage, on the general undergraduate course page. As an opposite, those users who arrived by organic search and those from residual regions and pages tend not to come back to visit the website once again.

This work highlighted that data mining can help identifying patterns in the user behaviour of a HEI website. Besides, there are several possibilities to be extended in terms of scope and focus on the experience of the HEI applicant, the data sources and the objectives of insight generation. As a contribution, the content of this work produced two papers presented at the conferences International Conference of Marketing and Technologies 2023 and at the International Conference of Knowledge Discovery and Information Retrieval 2023.

In summary, this work added knowledge on the application on a case study of a Portuguese Higher Education Institution of the concepts of Digital Marketing, Customer Experience, Web Analytics and Data Mining applying the CRISP-DM methodology.

The limitations of these study were the number of data sources, which was the only one from the website's visitor behaviour at the time of the work.

The outcomes of this study leave several future research possibilities to go further on the topic such as making comparison studies with other universities, planning and implementing new data sources for institutions as the one in this study and investigating the available data to answer more questions regarding the goals set in this study of for new goals.


Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

# REFERENCES


Analytics Help. (2023, July 12). *[GA4] Automatically collected events*. https://support.google.com/analytics/answer/9234069?sjid=10976514797204882090-EU#first_visit

Analytics Help. (2023, July 12). *[GA4] Analytics dimensions and metrics*. https://support.google.com/analytics/answer/9143382?sjid=10976514797204882090-EU#engaged-sessions&zippy=%2Csession

Bailyn, E. (2022, March 16). *Average Session Duration: Industry Benchmarks*. FirstPageSage. https://firstpagesage.com/seo-blog/average-session-duration-by-industry/

Basha, A. (2019). A study on the effective digital marketing strategy in education sector at Bangalore city. IJRAR-International Journal of Research and Analytical Reviews (IJRAR), 6(1), 161-169.

Bateman, T. S. (2021). Using Academic Social Networks to Enhance the Student Experience in Online Education. *Online Learning*, *25*(4), 296-323.

del Rocío Bonilla, M., Perea, E., del Olmo, J. L., & Corrons, A. (2020). Insights into user engagement on social media. Case study of a higher education institution. *Journal of Marketing for Higher Education*, *30*(1), 145-160.

Donovan, Robert J., Garry Egger, and Mark Francas. "TARPARE: a method for selecting target audiences for public health interventions." Australian and New Zealand journal of public health 23.3 (1999): 280-284.

Forghani, E., Sheikh, R., Hosseini, S. M. H., & Sana, S. S. (2022). The impact of digital marketing strategies on customer’s buying behavior in online shopping using the rough set theory. International journal of system assurance engineering and management, 1-16.

Harbi, A. M., & Ali, M. M. (2022). Adoption of Digital Marketing in Educational Institutions: A Critical Literature Review. *IJCSNS*, *463*.

Holiday, R. (2018). *Growth Hacker Marketing: A Primer on the future of PR, marketing, and advertising*. New York, NY: Portfolio/Penguin.

Hotz, N. (2023, January 19). What is CRISP DM? Data Science Process Alliance.

Kalbach, J. (2021). *Mapping experiences: A complete guide to customer alignment through journeys, blueprints, and diagrams*. Beijing: O’Reilly.

Kusumawati, A. (2019). Impact of digital marketing on student decision-making process of higher education institution: A case of Indonesia. *Journal of E-Learning and Higher Education*, *1*(1), 1-11.

Mandalapu, V., & Gong, J. (2019). Studying Factors Influencing the Prediction of Student STEM and Non-STEM Career Choice. In M. Desmarais, C. F. Lynch, A. Merceron, & R. Nkambou (Eds.), *The 12th International Conference on Educational Data Mining* (pp. 607–610).

Müller, M., & Aleksa Varga, M. (2020). Digital competences of teachers and associates at higher educational institutions in the Republic of Croatia. *Informatologia*, *53*(1-2), 8-23.

Nuriadi, N. (2021). The Effectiveness Of Application Of Marketing Strategies In Private Higher Education. *AKADEMIK: Jurnal Mahasiswa Humanis*, *1*(3), 104-113.

Nurtirtawaty, I. G. A. S., Murni, N. G. N. S., Bagiastuti, N. K., & Ruki, M. (2021). Digital marketing strategy through mobile application to increase room sales At Ibis



Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Styles Bali Legian Hotel. *Journal of Applied Sciences in Travel and Hospitality*, *4*(2), 93-100.

Oré Calixto, S. (2021). The effect of digital marketing on customer relationship management in the education sector: Peruvian case.

Osman, A. S. (2019). Data mining techniques.

Palomino, F., Paz, F., & Moquillaza, A. (2021, July). Web Analytics for User Experience: A Systematic Literature Review. In *International Conference on Human-Computer Interaction* (pp. 312-326). Springer, Cham.

Peling, I. B. A., Arnawan, I. N., Arthawan, I. P. A., & Janardana, I. G. N. (2017). Implementation of Data Mining To Predict Period of Students Study Using Naive Bayes Algorithm. *International Journal of Engineering and Emerging Technology*, *2*(1), 53. https://doi.org/10.24843/IJEET.2017.v02.i01.p11

Pinto, V.; Belfo, F.; Pedrosa, I. and Valgimigli, L. (2023a). Undergraduate candidate experience and engagement: insights from a case using CRISP-DM and Machine Learning.  In Proceedings of the 2023 International Conference on Marketing and Technologies.

Pinto, V.; Belfo, F.; Pedrosa, I. and Valgimigli, L. (2023b). Machine Learning in Customer-Centric Web Design: The Website of a Portuguese Higher Education Institution.  In Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 1: KDIR, ISBN 978-989-758-671-2, ISSN 2184-3228, pages 387-394.

Rajkumar, S. G., Joseph, D. C. S., & Sudhakar, D. J. C. (2021). Digital Marketing Communication Strategies and Its Impact On Student Higher Education Decision Making Process–A Review Of Relevant Academic Literature. *Psychology and education*, *13*.

RapidMiner. (2023, July 13). *Quality Measures*. https://docs.rapidminer.com/latest/studio/operators/cleansing/quality_measures.html

Saputro, B., Ma'mun, S., Budi, I., Santoso, A. B., & Putra, P. K. (2021). Customer churn factors detection in Indonesian postpaid telecommunication services as a supporting medium for preventing waste of IT components. *IOP Conference Series: Earth and Environmental Science*, *700*(1), 012015. https://doi.org/10.1088/1755-1315/700/1/012015

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, 526-534.

SimilarWeb. (2022, November). *Top Education Websites Ranking in Portugal in November 2022*. SimilarWeb. https://www.similarweb.com/top-websites/portugal/science-and-education/education/

Sheth, J. N., Newman, B. I., & Gross, B. L. (1991). Why we buy what we buy: A theory of consumption values. *Journal of business research*, *22*(2), 159-170.

Stone, M., & Woodcock, N. (2013). Social intelligence in customer engagement. Journal of Strategic Marketing, 21(5), 394-401.

Tiago, M. T. P. M. B., & Veríssimo, J. M. C. (2014). Digital marketing and social media: Why bother?. *Business horizons*, *57*(6), 703-708.

Uyanık, G. K., & Güler, N. (2013). A Study on Multiple Linear Regression Analysis. *Procedia - Social and Behavioral Sciences*, *106*, 234–240. https://doi.org/10.1016/j.sbspro.2013.12.027



Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Yassir, A. H., Mohammed, A. A.-J., Alkhazraji, A. A.-J., & Hameed, M. E. (2020). Sentimental classification analysis of polarity multi-view textual data using data mining techniques Data mining Data science Polarity text data Sentiment analysis Text data mining. *International Journal of Electrical and Computer Engineering*.

Zhu, W., Zeng, N., & Wang, N. (2010). Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations. *NESUG Proceedings: Health Care and Life Sciences, Baltimore*.



Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt