Data mining and predictive analytics for e-Tourism

Authors:

Antonio, Nuno [1,2], Almeida, Ana de [1,3,4], Nunes, Luis [1,4,5]

ISCTE-IUL[1], Nova IMS[2], CISUC[3], ISTAR-IUL[4], Instituto de Telecomunicações[5]

Abstract

Computers and devices, today omnipresent in daily life, foster the generation of vast amounts of data. Turning data into information and knowledge is the core of data mining and predictive analytics. Data mining uses machine learning, statistics, data visualization, databases, and other computer science methods to find patterns in data and extract knowledge from information. While data mining is usually associated with causal-explanatory statistical modeling, predictive analytics is associated with empirical prediction modeling, including the assessment of the prediction's quality. This chapter will present the basic concepts, fundamentals, and definitions of data mining and predictive analytics, including the type of problems for which it can be applied to and the most common methods and techniques employed. Some areas of e-Tourism that could benefit from the application of these methods and techniques will also be discussed. The main goal is that of providing the readers, even those unfamiliar with this topic, a general overview of the key concepts and potential applications of data mining and predictive analytics and entice the readers to apply these concepts in e-Tourism research projects. As such, an explanation of the life cycle of Data Mining and Predictive Analytics projects is offered, describing the tasks that compose the most widely employed process model, both for industry and academia: CRoss-Industry Standard Process for Data Mining, CRISP-DM.

Keywords

Database mining; Data visualization; Knowledge discovery; Machine learning; Predictive Analytics; Predictive Modeling;

1. Introduction

Computers and devices that today are omnipresent in daily life foster the generation of vast amounts of data. With the increase in the quantity of data, the capability to conveniently or automatically collect and process it to find patterns and extract knowledge also increases (Witten et al. 2011; Han et al. 2012). The process of identifying original, useful, and understandable patterns from data is known as Knowledge Discovery in Databases (KDD). Data Mining resides at the core of the KDD process and involves the use of different algorithms to build models of patterns in data (such as trends, associations, or affinities), or predictive models (Maimon and Rokach 2010; Delen and Demirkan 2013; Zaki and Meira 2014; Malik et al. 2018). Above all, patterns have to be “interesting”. To be considered interesting, patterns should be “(1) easily understood by humans, (2) valid on new or test data with some degree of certainty, (3) potentially useful, and (4) novel. A pattern is also interesting if it validates a hypothesis that the user sought to confirm. An “interesting pattern represents knowledge”(Han et al. 2012, p. 21). Although not consensual, it is not surprising that several researchers acknowledge that the use of Data Mining and Big Data is inducing a paradigm shift from scientific research to data-driven research (Strasser 2012; Kitchin 2014; Mazzocchi 2015). A shift that runs counter the deductive approach in current science: (1) Big Data can capture the whole domain; (2) there is no need for a priori theory or hypothesis; (3) data analytics methods can give a

non-biased view of patterns in data; (4) results can be interpreted by someone who understands statistics or data visualizations (even with limited domain knowledge) (Kitchin 2014).
Data mining tasks are described in two categories: descriptive and predictive. While descriptive tasks characterize data properties, predictive tasks perform induction in order to make estimations and predictions (Han et al. 2012). From a taxonomical point of view, Analytics is divided into three main categories (Delen and Demirkan 2013):

- Descriptive Analytics: uses data and standard reporting/aggregations to answers the questions “What happened?” or “What is happening?”;
- Predictive Analytics: uses data and algorithms to build mathematical models to discover explanatory and predictive patterns that can answer the questions "What will happen?" or "Why will it happen?”;
- Prescriptive Analytics: uses data and algorithms to build mathematical models to determine high-value "courses-of-actions" or "what-if scenarios". The objective is to answer the questions "What should I do?" or "Why should I do it?".

Together with Text Mining, Web and media mining, or statistical time series forecasting, Data Mining is considered one of the main enablers of Predictive Analytics (Delen and Demirkan 2013). Probably because of this, Data Mining and Predictive Analytics are either used interchangeably (e.g., Malik et al. (2018)) or used to identify the type of task to be undertaken. While Data Mining is employed for causal-explanatory statistical modeling (descriptive tasks), Predictive Analytics is used to describe predictions and estimations of future outcomes (predictive tasks) (e.g., Larose and Larose (2015)).

Data Mining can be applied to almost any kind of data as long as the data are meaningful for the required application, which describes mostly all types of data: database data, data warehouse data, transactional data, hypertext and multimedia data (text, image, video, or audio), graph and network data (e.g., social and information networks), among others (Han et al. 2012). However, when modeling is intended, most data mining techniques requires the data to be transformed into a two-dimensional dataset, where rows represent the unit of analysis (e.g., traveler, customer, online review), and a column represents the measure of an attribute (e.g., year of birth, nationality, review rating) (Zaki and Meira 2014; Hastie et al. 2017). Take as an example, a Data Mining project to analyze online reviews' ratings where the dataset should be structured like the one shown in Table 1. Depending on the field of study, the attributes (columns) in the dataset can assume different names. The column with the dependent variable, the variable that is the focus of the problem, is also known as a response variable, target or label (“Rating” in Table 1 example). The remaining columns are usually known as independent variables, explanatory variables, predictors, or features. Variables can be classified into two main types (Zaki and Meira 2014): (1) Numeric – real or an integer-valued and assuming a finite or infinite set of values, respectively classified as discrete or continuous; (2) Categorical – a set-valued domain (e.g., gender or education levels). Categorical attributes can also be classified as nominal (when no order is assumed – for example, gender) or ordinal (when the values are ordered in the form of a ranking – for example, education levels). The process to carry on a Data Mining and Predictive Analytics project, including the preparation of the modeling dataset and its attributes, is presented in section 2.3.

Rating	User ID	User Location	User Age	Attraction	Comment
5	John Doe 1	London, UK	30-40	London Eye	Excellent, bla, …
4	Mary Doe 1	NY, USA	50-60	Tower of London	Very good, bla, …
5	John Doe 2	Manchester, UK	30-40	Windsor Castle	Excellent, bla, …
3	Mary Doe 2	Chicago, USA	20-30	Tower of London	Not so good, bla, …
…	…	…	…	…	…

Table 1. Example of a dataset structure (online reviews ratings)

As illustrated in Figure 1, Data Mining incorporates techniques from different domains, including (Han et al. 2012; Zaki and Meira 2014):

- Statistics: exploratory study, analysis, interpretation or explanation of data.
- Data visualization: a visual representation of data.
- Information retrieval: search of documents or information in documents.
- Data warehouse and database systems: creation, maintenance, and use of databases, including the integration of data from multiple sources to build systematic data analysis capabilities (data warehouses).
- Machine learning: investigates how computers can learn from data.

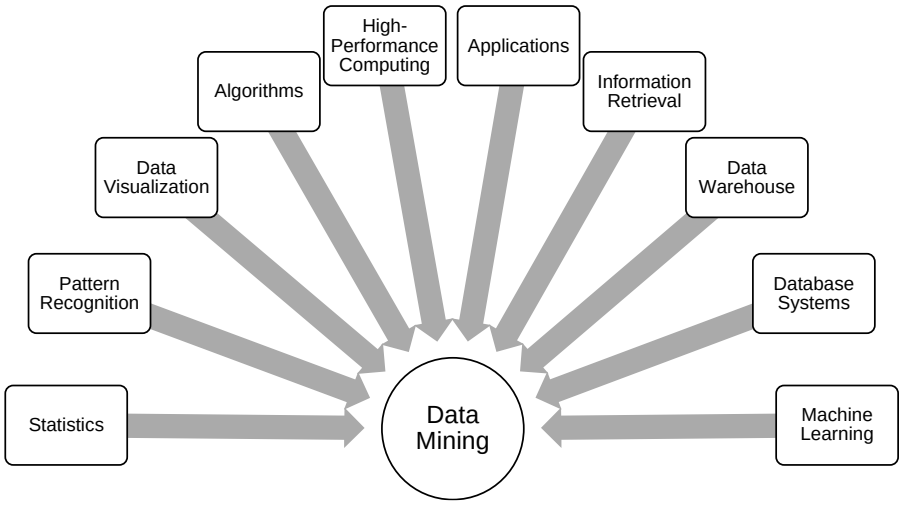

Figure 1. Techniques adapted from the different domains by Data Mining (adapted from Han et al. (2012)).

Machine Learning plays an essential role in Data Mining (Witten et al. 2011; Han et al. 2012). Figure 2 illustrates some of the types of Machine Learning problems that are highly related to Data Mining (Han et al. 2012; Hastie et al. 2017):

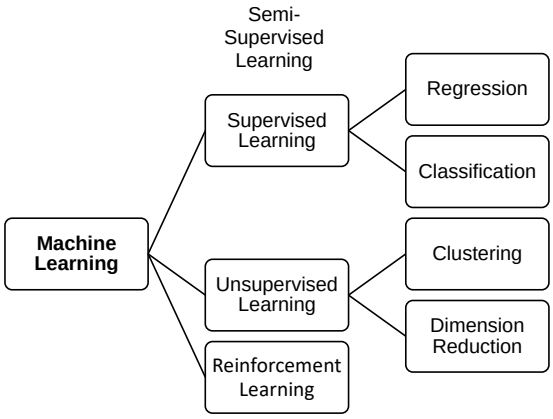

Figure 2. Machine Learning's types of problems

- Supervised learning: uses labeled input attributes to predict an outcome/target. Supervised learning problems are commonly divided into:
    - Classification problems: when the target is discrete. The goal is to identify to which of the possible classes the dataset observation belongs to (e.g., classify travelers by

predefined segments based on the time to service they book their trips and the length of the stay like in the visual illustration shown in Figure 3).
    - Regression problems: when the target is a continuous measurement (e.g., predict the average occupation rate of a hotel or the average amount spent per tourist in a destination).
- Unsupervised learning: when the input attributes are not labeled, and there is no target. The goal of unsupervised learning is to explain the associations and patterns between input attributes, and it is divided into:
    - Clustering: identifies clusters of input attributes (e.g., identifies customer segments);
    - Dimension reduction: "simplify" data by aggregating variables with similar contribution to the dataset variance.
- Semi-supervised learning: makes use of non-labeled input attributes to gain more understanding of the population.
- Reinforcement learning: when the input attributes are not labeled or labels are not defined, and the model learns (improves performance) based on a rewarding process.

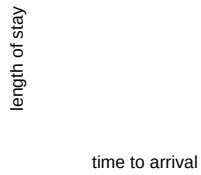

Figure 3. Classification model visualization example (identify customers' segment based on two attributes)

Although Information Communication Technologies (ICT's) have been transforming tourism globally since the 1980s (Buhalis and Law 2008) until 2007 literature related to the application of Data Mining and Machine Learning in tourism and travel research was still scarce (Delen and Sirakaya 2006; Law et al. 2007). However, this scarcity began to reverse from thereon and research on multiple topics has been published since then, like for trend analysis/forecasting (Wu et al. 2010; Claveria et al. 2015; Moro and Rita 2016; Höpken et al. 2020), travel planning/recommendation (Hsu et al. 2010; Zhao and Ji 2013), understanding travelers' preferences/personalization (Li et al. 2015; Chang et al. 2016; Shapoval et al. 2017; Antonio et al. 2018), understanding travel patterns (Chen et al. 2018; Hu et al. 2019), analysis of customers' profitability (Pei 2013), analysis of destination competitiveness (Srivihok and Intrapairot 2016), impact on Customer Relationship Management (Xie and Tang 2009), predicting and understanding hotel cancellation drivers (Falk and Vieru 2018; Antonio et al. 2019), among others.

Bach et al. (2013) performed a keyword analysis of the literature published between 1995 and 2013 and identified 88 peer-reviewed articles for the application of Data Mining in tourism. In these, the authors identified six core areas of application: forecasting, personalization, tourism management, tourism systems (e.g., recommender systems), multi-agent systems (e.g., swarm optimization), and Machine Learning based applications.

More recently, Mariani et al. (2018), in a literature review about research published between 2000 and 2016 on Business Intelligence and Big Data in hospitality and tourism, based on a search using, among others, the keywords "Business Intelligence", "Data Mining", and "Data Warehouse", found that there was a wide distribution of publications along those years, presenting a linear and relevant growth over time. The authors concluded on the existence of an upsurge in the number of research publications in hospitality and tourism literature that apply analytical techniques to large quantities

of data. However, they also concluded that research is somewhat fragmented in scope, limited in methodologies, and inclusively, shows some gaps.

In fact, despite the growing interest in the application of Data Mining in tourism research, reflected by the increasing number of publications and topics covered, there is still much to explore. The potential of data-driven research is enormous.

2. Patterns, applications, and processes

In this section, we will present the kind of patterns that Data Mining and Predictive Analytics can expose to help explain and predict the tourism phenomena. We will also describe possible applications and the process to conduct Data Mining and Predictive Analytics based projects.

2.1. Pattern types

As illustrated in Figure 4, five primary types of patterns can be mined (Han et al. 2012; Zaki and Meira 2014), that is next described as:

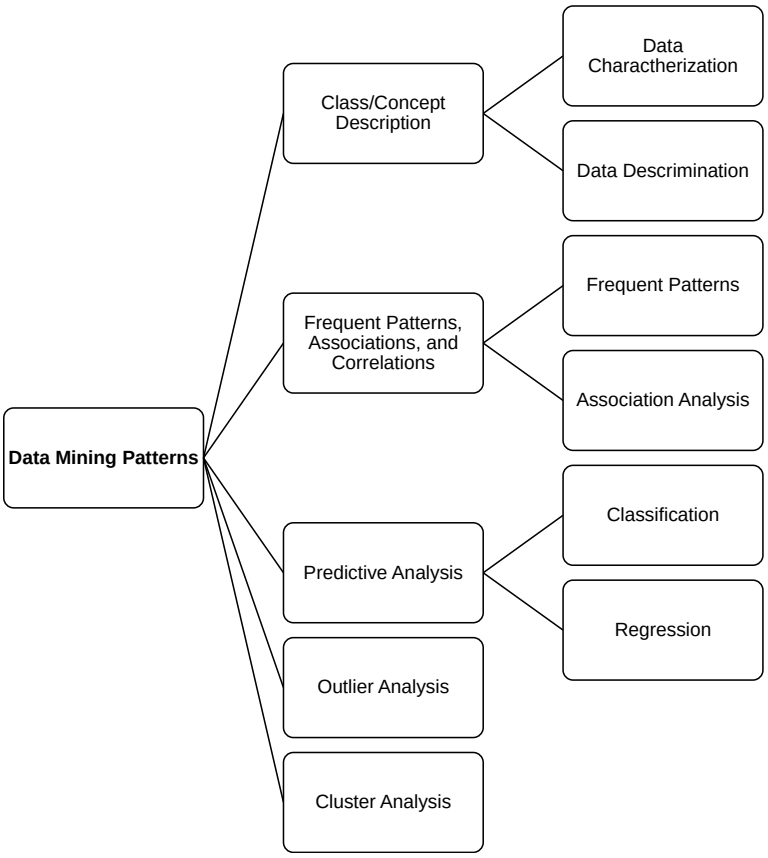

Figure 4. Data Mining types of patterns

- Class/concept description: description of a class or concept that can be derived from the summarization of the class under study (often called target or outcome), which is data characterization, or the comparison of the target class with other classes, which is data discrimination, or both. While the outputs of data characterization are charts (e.g., pie, bar, or columns), single or multidimensional tables, or in rules form, data discrimination outputs are mostly done in the form of rules. An example of these type of patterns could be the study in a hotel company of segment distribution by nationality and distribution channel.
- Frequent patterns, associations, and correlations: frequent patterns are patterns that frequently occur in data, such as frequent itemsets, frequent subsequences (known as sequential patterns), and frequent substructures. Frequent itemsets usually refer to a set of items that often appear together in a transactional dataset. Frequent subsequences can refer

to the sequential order, obtained for purchase history, that a traveler tends to buy the traveling services (e.g., first airline ticket, followed by a hotel, then transfer). A frequent substructure can refer to the combination of frequent itemsets and frequent subsequences with other structures (e.g., graphs or trees). These patterns are usually designated as "Association rules" in Machine Learning.

- Predictive analysis: the process of finding a model (function) that predicts categorical labels (classification) or that predicts continuous-valued labels (regression). In classification or regression problems, the modeling dataset must be fully labeled, i.e., one of the attributes (columns) of the dataset must be the label. For example, to forecast hotel room occupation demand the dataset should be comprised of attributes of monthly demand per country, distribution channel, market segment, room types, among others, and also an attribute as the label, in the case, the number of rooms occupied.
- Outlier analysis: the identification of observations in a dataset that do not comply with the general pattern or model of the data (also known as anomaly mining or anomaly detection). Outlier analysis is used, for example, in fraud detection or for the identification of customers with high Customer Lifetime Value (CLV).
- Cluster analysis: contrarily to supervised learning, with labeled-data, unsupervised learning is employed to analyze non-labeled data. One example of unsupervised learning methods is cluster analysis. As illustrated in Figure 5, cluster analysis could be used to identify groups in data (e.g., customers segments).

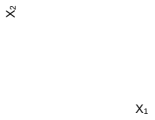

Figure 5. Clustering visualization example (analyzing travelers per amount spent and time to service booking)

2.2. How to choose the technique/tool that best suits the problem type

For most Data Mining problems, multiple tools/techniques/algorithms can be applied. Each one has its particular conditions for application, as well as consequences and trade-offs (Maimon and Rokach 2010; Han et al. 2012). In this sub-section, we will introduce some of the techniques/algorithms employed in the different type of problems, but we will not delve into the detail of each one, as that would be out of the scope of this chapter. Details of techniques/algorithms can be found in Data Mining reference literature, such as Hastie et al. (2001), Maimon and Rokach (2010), Han et al. (2012), or Zaki and Meira (2014). To better illustrate which tools/techniques/algorithms to select according to the problem, we present the next flowchart diagram (Figure 6). While not exhaustive, this visual aid attempts to describe the most popular tools and techniques. The process for selecting the best one for a particular Data Mining project will be discussed in section 2.3.

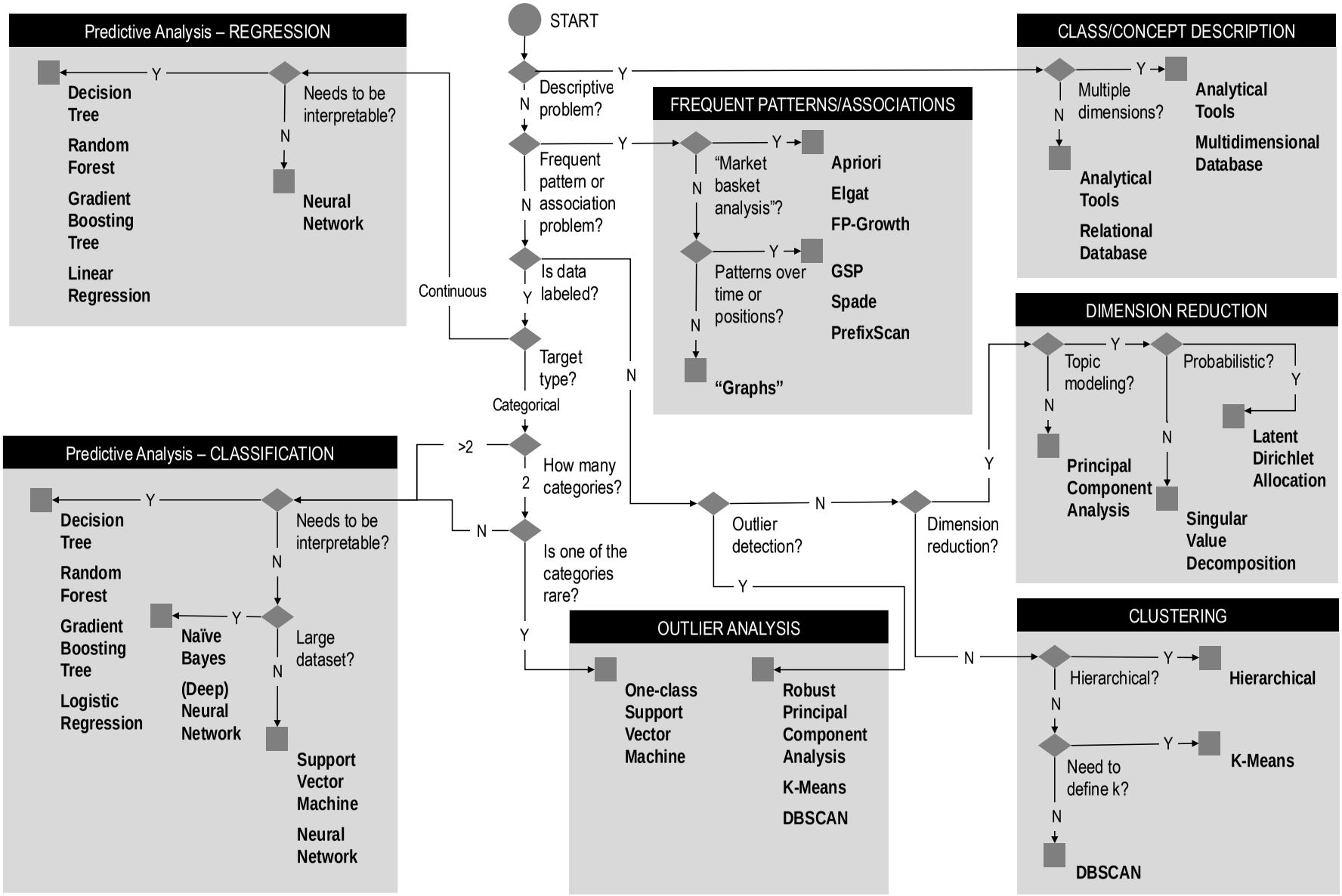

Figure 6. Algorithm/tool selection diagram

2.2.1. Class/concept description

The type of problems for class or concept description are frequently descriptive problems, that is, problems that rely on statistics summarization, statistical tests, and aggregations, whose output is presented in the form of visualizations, tables, crosstabs, or rules.

When made from a single dataset, summarizations statistics and aggregations can be implemented (as most of the techniques described in this section) in free, non-commercial, programming languages such as Python, R, or tools such as SAS Miner or Weka. However, if data is multidimensional, it should be stored into a relational or a dimensional database. However, this is not necessarily implemented in a Database Management System (DMBS), as many of the actual analytical tools, including spreadsheets, support the creation of dimensional databases models. Relational databases also referred to as Online Transaction Processing (OLTP) systems, are user-oriented databases, employed for transaction and query processing. Dimensional databases, also referred to as Online Analytical Processing (OLAP) systems, are decision-making oriented and used for data analysis. OLAP systems allow data to be modeled and viewed in multiple dimensions – usually designated as a "cube" (Han et al. 2012). In case the goal involves analysis of a multidimensional problem (such as the study of the impact of different dimensions and factors, like the weather forecast, currency exchange, or local events on travelers' demand) then a multidimensional database should be used. OLAP databases allow the creation of measures based on aggregation functions, such as the count of frequencies, minimum values, maximum values, mean, median, rank order, among many others. The measures can be consulted through simple reporting or explored through the four main types of OLAP operations (see examples in Figure 7):

- Roll-up: performs aggregation on a data cube, either climbing up a concept hierarchy in a dimension or by dimension reduction.
- Drill-down: drill-down is the reverse of roll-up. It allows navigation from an aggregated level of data to a more detailed level.
- Slice and dice: while "slice" create a sub-cube by performing a selection in one of the dimensions, "dice" creates a sub-cube by performing a selection in two or more dimensions. For example: (month="January") AND (region="NY");
- Pivot (rotate): visualization operation which rotates the data axes in view to offer an alternative view on data (from another point of view). For example, instead of looking on the average length of stay travelers, per region and year (lines), per month (column), change the year to columns and the month to the lines of the report.

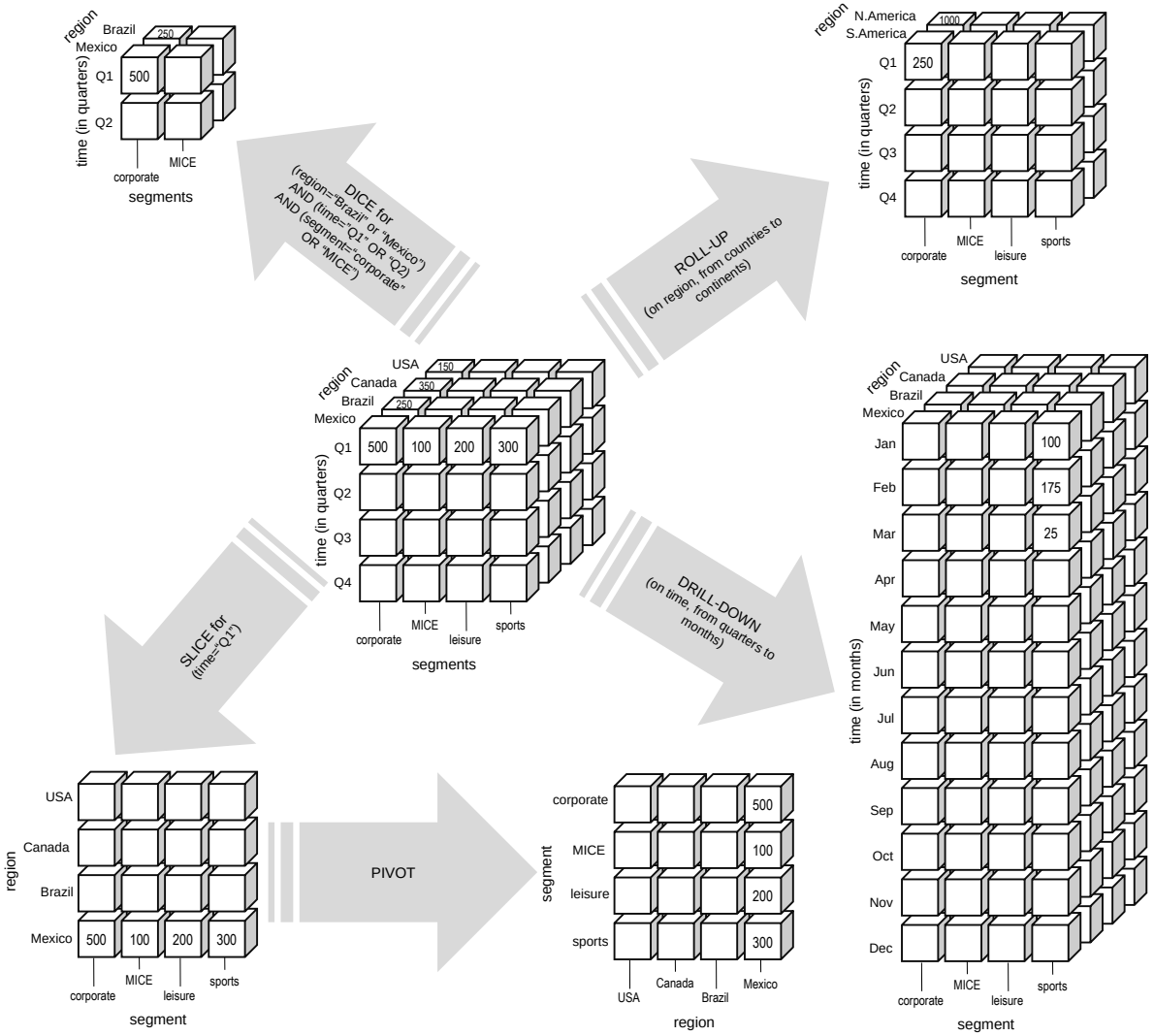

Figure 7. Examples of main OLAP operations (adapted from Han et al. (2012))

2.2.2. Frequent patterns, associations, and correlations

Frequent pattern mining investigates recurring relationships in a data set, like the co-occurrence of two or more objects of interest (Han et al. 2012; Larose and Larose 2015). The more commonly known application of frequent pattern mining is the “Market basket analysis”, that is, to identify sets of items that are purchased together (for example, in supermarket transactions). However, "Market basket analysis" can be used not only to identify customers purchase habits but also in e-Tourism applications. Examples of e-Tourism applications could be the identification of which traveling services were purchased together in an online traveling website, understanding web pages visiting patterns of an official destination website or understanding visitors' patterns in a trip by analyzing the online reviews posted by the travelers. The most common techniques to identify co-occurrence patterns are the Apriori algorithm, the Eclat algorithm, and the FP-Growth algorithm. Although the first is the most employed and the most known, the other algorithms were designed to overcome some of the limitations of the Apriori algorithm, such as the high computational requirements for processing large datasets (Han et al. 2012; Larose and Larose 2015). One example of the "Market basket analysis" in e-Tourism research is the work of Liao et al. (2010), where the authors combined the application of the Apriori algorithm with clustering to propose suggestions and solutions for tourism product development.

Also interesting in e-Tourism research is sequence patterns mining. Sequence mining allows the discovery of patterns across time or positions in a dataset (Larose and Larose 2015). Sequence patterns could be used to identify, across time, what travelers say in online reviews about a destination or a hotel brand, or travelers purchase patterns in hotels across time, or travelers' destination type selection through time, or, as in Bermingham and Lee (2014) and Cai et al. (2014), explore travelers' trajectories. There are several techniques to identify sequence patterns, like the GSP, Spade, or PrefixScan algorithms.

Graph mining is one other interesting pattern mining tool. With the ubiquitous presence of social networks, graph data has grown in importance. Graph mining aims to find interesting subgraphs in data. For those not familiar with the term, a graph is a structure that represents a set of objects that may be interconnected. Objects are usually named vertices or nodes, and the links between them are called edges or links. One example of a graph is the representation of "friendship" in a social network, with users being the vertices and the "friendship" between users being the edges (Figure 8). Graphs analysis can be used to study and measure service quality based on online reviews (Li et al. 2010), to understand travelers' movements patterns in a destination (Hu et al. 2019), or to identify social media influential users and predict their network impact (Francalanci and Hussain 2016). The more common graph pattern mining is the mining of frequent subgraphs from a database of graphs, which is usually done with the GSpan algorithm (Han et al. 2012; Zaki and Meira 2014). Other patterns such as closed graphs, coherent graphs, or dense graphs can also be mined.

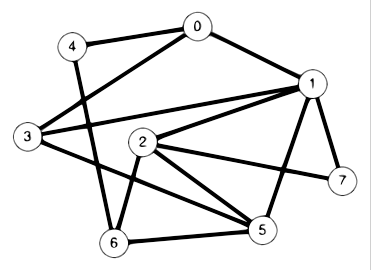

Figure 8. Example of a graph network representation

2.2.3. Predictive analysis

As mentioned in section 2.1, the objective of predictive analysis is to build a mathematical function (model) that, based on past events, can estimate future events. Conversely, sometimes, more than estimate or forecast future events researchers and modelers seek to uncover the causes for the events to happen. Uncovering the causes should be achieved not only by statistic modeling but also implies predictive modeling (Shmueli 2010). However, high-performance machine learning algorithms usually generate black-box models, that is, highly complex mathematical functions, making interpretation and understanding of the causal mechanisms behind the prediction virtually impossible. An example of these algorithms is artificial neural networks (ANN). ANNs are increasingly being employed in e-Tourism research (Moro and Rita 2016). ANNs are inspired by the biological neural networks of the human brain. ANNs are a collection of nodes (neurons), connected by edges (Figure 9). ANN's consist in an input layer (a_1 in Figure 9) with an entry for each input variable, an output layer ($a_4$in Figure 9) and can have one or more hidden layers ($a_2$and a_3 in Figure 9). The output is calculated by a non-linear function that sums the nodes' inputs multiplied by the edges' weights ($w_n$in Figure 9 example).

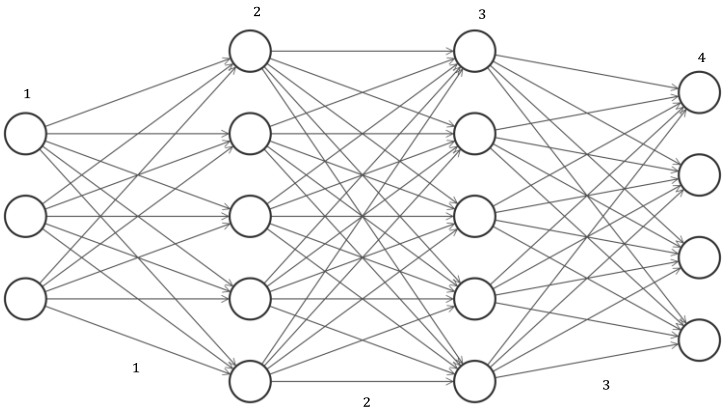

Figure 9. Example of a neural network (one input layer, two hidden layers, and one output layer)

Other algorithms, such as decision-tree based algorithms, are quite straightforward to interpret by humans, as shown by the example in Figure 10 that presents a model built for the well-know “play golf” dataset, where the label is a YES or NO answer to the “Play golf” questions. The predictors are weather-related variables: "Outlook", "Humidity", and "Windy". It is relatively easy to understand the model built from the 14 observations of the dataset. Besides allowing a certain degree of interpretation, decision-tree based algorithms also have the advantage of automatically handling missing data, incorporate the treatment of outliers, inherently detect variable interactions, and not being affected by variable skewness.

Additionally, decision-tree based algorithms have the advantage of being nonparametric in the sense that no statistical distribution assumptions are made about the explanatory variables and the label. Nevertheless, these algorithms also present some disadvantages, namely a tendency to overfit, i.e., perform well on training data but do not generalize well for unknown data. An optimized version consists in employing ensembles to overcome the overfitting issue. Ensemble methods combine the results of multiple trees into one model. The most known ones, Random Forest or Gradient Boosting Tree usually show excellent performance while maintaining a certain level of interpretability.

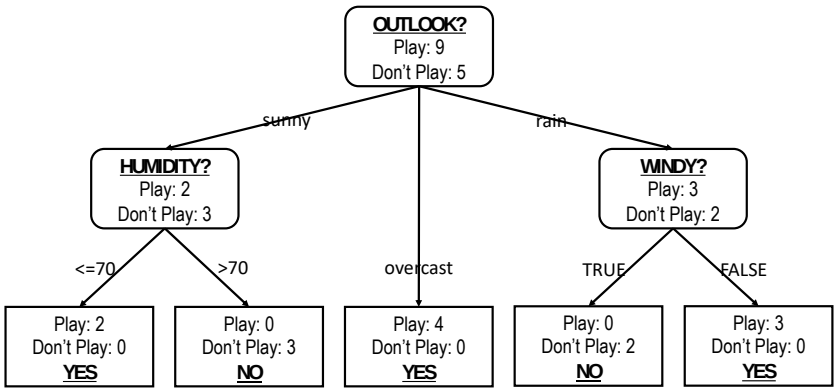

Figure 10. Decision-tree example (Play golf dataset: 22 weather condition observations)

On the one hand, has illustrated in Figure 6, not all algorithms facilitate model interpretability. Some of the algorithms generate models with a certain degree of interpretation, for both regression and classification problems. This interpretability is clear with Decision Tree, Random Forest, Gradient Boosting Tree, or Linear Regression for regression problems. A certain level of interpretation is also achievable for classification problems when using Decision Tree, Random Forest, Gradient Boosting Tree, or Logistic Regression.

On the other hand, high-performance algorithms, such as Neural Network or Support Vector Machine based algorithms, generate less interpretable models. It is up to the modeler to select the "right" algorithm, depending on the type of estimation problem and whether an understanding of what are the causal mechanisms behind the estimation is needed. For example, Falk and Vieru (2018) and Antonio et al. (2019), when predicting hotel bookings' cancellations outcome (cancel or not cancel, thus, a classification problem), opted for more interpretable algorithms which enable, not only to predict the cancellation outcome, but also to understand cancellation drivers. Another example of a regression problem where interpretability was not sought for is the research of Claveria et al. (2015) that employed different types of ANNs to forecast tourist demand.

2.2.4. Cluster analysis

Clustering (Figure 5) is a method for the partition of multiple entities into groups so that, within the same group, entities share a certain degree of similarity, but are ideally very dissimilar to the entities in the other groups (Han et al. 2012). There are several types of clustering methods: (1) Partitioning, that create completely separated partitions (groups); (2) Hierarchical, that build a hierarchical decomposition of the dataset, identifying subgroups inside groups; (3) Density-based , which differs from other techniques by enabling the definition of non-spherical-shaped clusters (hierarchical clusters); (4) Grid-based where the entities are quantized into a finite number of cells (grid).

The most employed clustering algorithm is K-Means, a partitioning method. After the modeler's definition of the number k of clusters for the partition of the dataset, clusters' centroids are redefined at every iteration, based on the average distance of the data points to the cluster centroid. A typical application of this algorithm in e-Tourism has to do with travelers' segmentation (Pesonen et al. 2011, Chen et al. 2014). One other popular clustering algorithm is DBSCAN. Instead of using the distance between data points, DBSCAN uses the local density of points to define the groups. Though slower than K-Means and requiring more parameters to be configured, DBSCAN has the advantage of not requiring a pre-definition of the number of groups (k).

2.2.5. Outlier analysis

An outlier (or anomaly) is a data point that deviates significantly from the dataset's remaining points (for an illustration see Figure 11). Thus, the objective of an outlier analysis is the discovery of unusual data points in the dataset. These data can occur due to human errors (data entry errors), experimental errors (errors related to data extraction or preparation), processing errors (data manipulation errors), sampling errors (incorrect sampling or data sources selection), or novelty errors (indicating, for example, new trends).

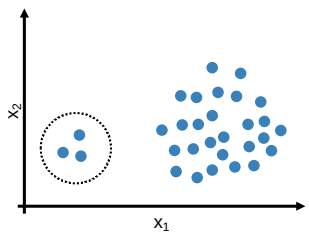

Figure 11. Example of outliers in a dataset (points inside the circle)

As in other types of Data Mining patterns, there are several techniques/algorithms for performing outlier analysis. If the dataset is labeled, the most common detection algorithm is a Machine Learning based algorithm, the One-Class Support Vector Machine, viewing the problem as a classification problem. If the dataset is not labeled, then a clustering algorithm such as K-Means or

DBSCAN is usually employed to identify data points that are further away from the cluster centroids. Another method used for unlabeled data is a version of PCA known as Robust Principal Component Analysis.

Besides being used in almost every exploratory analysis to understand the data better, outlier analysis can also be the central methodological approach for a research project. In tourism research, outlier analysis can be used, for example, to identify the variety of existing accommodations in a given region (Sánchez-Martín et al. 2019) or to discover new tourist points of attraction from social media data (Halim et al. 2018).

2.2.6. Data preprocessing

Before applying any of the previous mentioned techniques, as described in the following section, data must be processed and transformed. From the many existing transformations' techniques, two are worth to mention: Dimension reduction and Topic Modeling.

Dimension reduction is the process of scaling down the number of predictor variables, generating a reduced representation of the dataset, yet preserving the essential properties of the full dataset (Han et al. 2012; Zaki and Meira 2014). High-dimensional datasets can be complex to analyze, require processing time and, for predictive modeling, can lead to overfitting models. There are several dimensionality reduction algorithms, but Principal Component Analysis (PCA) is the most well-known and one of the most used. The goal of PCA is to project the original dataset onto a low dimensional space preserving as much variance as possible. At the same time, PCA also helps to understand the correlation between a set of predictor variables. This understanding was used by Brida et al. (2011) and Muresan et al. (2016) to understand residents' perceptions of tourism.

Topic modeling is a Text Mining technique to discover topics that are addressed within a collection of texts. However, Topic Modeling, like clustering, is sometimes used as a data reduction technique. By assigning each document a topic, the variables used in the identification of the topics can be replaced by one variable only, the identified topic.Two usual algorithms that are used for topic modeling are Singular Value Decomposition and Latent Dirichlet Allocation. A well-known application of these techniques/algorithms is the use of Latent Dirichlet Allocation to extract the dimensions of traveler's satisfaction and dissatisfaction from online reviews (Rossetti et al. 2016; Guo et al. 2017).

2.3. ***How to conduct a Data Mining project***

Cross-Industry Standard Process Mining for Data Mining (CRISP-DM) (Chapman et al. 2000) is today the de facto standard process model employed for Data Mining projects, both in industry and academy. The reason for CRISP-DM's popularity seems to be related to its foundation in technical principles originated from practical and real-world experience on how modelers conduct data mining projects.

CRISP-DM provides a life cycle of a Data Mining project (Figure 12) that is composed of six phases, with multiple tasks in each phase. Phase sequencing is not rigid, and it is usual to move back and forth between several steps. The outcome of one phase or task defines the task or phase that is performed next. The arrows connecting each phase in Figure 12 illustrate the principal and most frequent interactions. Multiple iterations between different phases are usually necessary until a model can be deployed. The outer arrows in Figure 12 represent the cyclical nature of Data Mining projects. Projects do not end when models are deployed. Lessons acquired from the modeling process and, in particular, changes in the phenomena under study brought by the model deployment, are reincorporated for the model's continuous improvement.

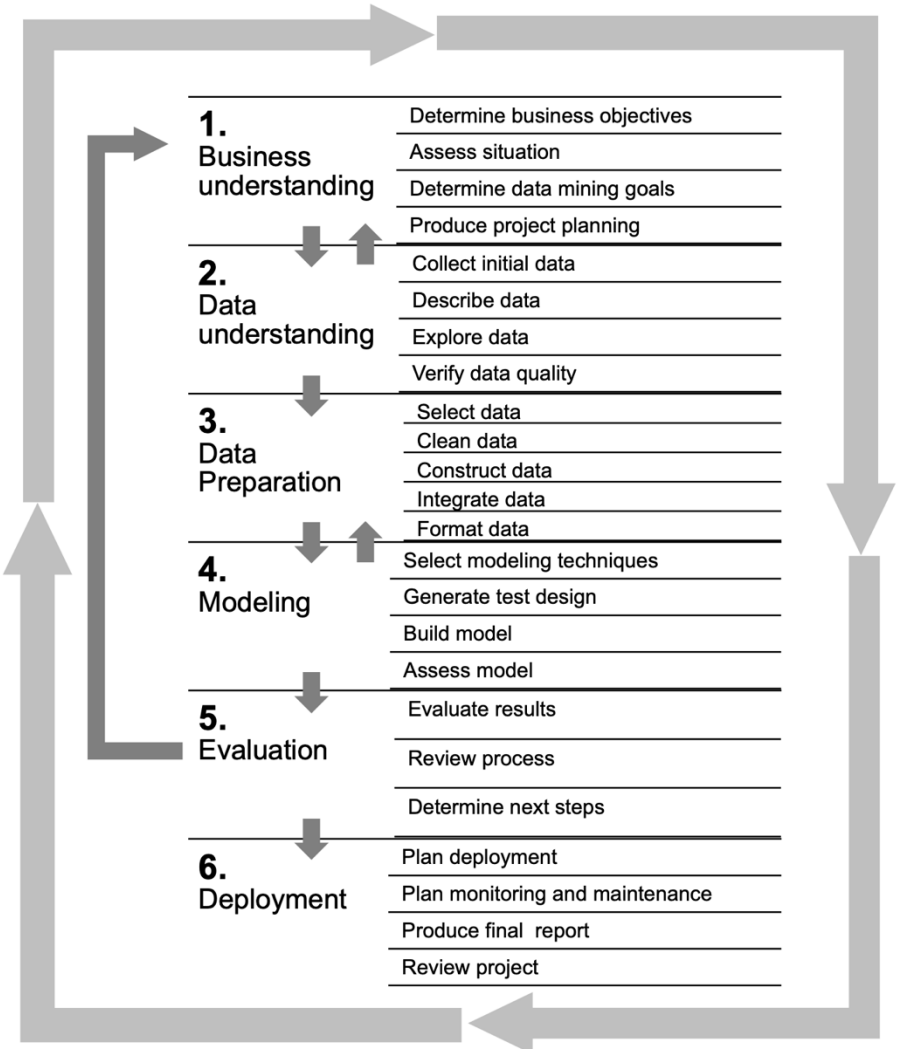

Figure 12. Phases and generic tasks of the CRISP-DM reference model

Without entering into a high level of detail, this section describes the generic tasks of each phase, including their outputs and main challenges:

1. Business understanding: the initial phase where project objectives and requirements are studied from the problem and domain perspective and converted into an analytics project, resulting in the definition of a plan to achieve the objectives. As Han et al. (2012) recognize "It's tempting to jump straight into mining" but, for a Data Mining project to be successful, special attention should be paid to the problem's goal and domain.
    a. Determine business objectives: understand the problem correctly, balancing objectives with constraints, accurately identify the questions to be answered, in order to avoid wasting a great deal of effort producing the right answers to the wrong questions.
    b. Assess situation: detailed identification of all available resources, constraints, assumptions, and other factors, including data, stakeholders, experts, computational resources, among others that could influence project design and goal achievement.
    c. Determine data mining goals: define the project goals and the criteria for a successful outcome in technical terms (e.g., a certain level of predictive accuracy);
    d. Project planning: describe in detail the plan to answer the objective questions and the project goal.

2. Data understanding: starts with the initial data collection and continues with the activities required to allow modelers to become acquainted with data, including identifying any patterns, tendencies, or anomalies.
    a. Data initial collection: attain the data identified in the project resources. Data collection could be done through SQL database queries, accessing files in data warehouses, downloading data from the internet through the use of websites' Application Programming Interfaces (API), websites' scraping, or any other required methods. For this task output, besides the obtained data, also the methods and the detailed description of how the data was obtained is important for the future.
    b. Data Description: analyze the general data details and generate a report with the basic properties of data (e.g., number of columns or number of observations) to evaluate if the data so far satisfy project requirements.
    c. Data exploration: statistics, reporting, and data visualization to study data in order to perform exploratory analysis, including distribution of key variables, correlations, results of simple aggregations, categorical variables classes frequencies, properties of significant sub-populations, among other statistical analyses.
    d. Data quality check: identify any quality issues in data (does data covers all cases?), discover errors and typos (e.g., different names are used to identify the same classes in categorical variables), missing values, among other issues. If issues arise, enumerate them and define possible solutions (knowledge in the domain of the problem is usually a key factor to overcome data quality problems).
3. Data preparation: all activities related to the creation of the final dataset (modeling dataset).
    a. Data selection: select the data (columns and rows) to be included in the modeling dataset based on criteria taking in consideration the goal of the project, data quality, and the resources available, including, among others, the computational power required.
    b. Data cleaning: improve data quality by, for instance, cleaning sub-sets of data, inserting of default values, or estimating missing values.
    c. Data construction: construct new observations from aggregations of observations or variables derived from other variables (known as “engineered features” in Machine Learning – e.g., a ratio built from two other variables). Engineered features are considered a critical success factor in predictive modeling due to information gain obtained from the association of multiple input variables. Sound feature engineering requires, not only technical knowledge, but also creativity, intuition, and domain knowledge (Antonio et al. 2019);
    d. Data integration: merge data from different sources or constructed data from the previous task, if any.
    e. Data Formatting: format data according to algorithm requirements (for example, the label must be the first column in the dataset, binary categorical variables classes are expressed as 0 and 1 or change the distribution of highly skewed variable with a mathematical function such as the $\log_{10}$).
4. Modeling: comprises all final activities related to the preparation of the dataset for modeling and the application of the chosen algorithm, including parameters’ calibration.
    a. Modeling techniques: select the technique/algorithm to use according to the problem (e.g., neural network or decision tree). If multiple techniques/algorithms must be applied, this task should be performed separately for each technique/algorithm.
    b. Test design generation: set up the tests to evaluate the model’s quality and validate it. For predictive modeling, this is in this task where the dataset is typically divided

into a training set (usually comprising between 60% to 80% of the data) and a test set (with the remaining data). While the training set is used to build the model, the model's quality is tested using the test set.

c. Modeling: using the modeling tool, create the model.
d. Model assessment: analysis of the model's performance according to technical performance and kind of model. For example, for classification models, measures such as Accuracy, Precision, Recall, F1Score, or Area Under the Curve (AUC) are commonly employed. In regression problems, measures such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) or Mean Squared Error (MSE) are used. For clustering, different measures are employed, such as Average Silhouette Width (ASW), p-Separation Index, p-Stability Index, among others. If models were built using different techniques/algorithms, make relative performance comparisons.

5. Evaluation: assessment of the model's performance according to the objectives initially determined.
    a. Evaluate results: while in the previous phase, results were evaluated from the technical perspective, now results are analyzed in terms of the project goal and questions. The tasks' output should be an assessment based on the success criteria.
    b. Review process: revise all the process to identify any undone activities and those that should be repeated.
    c. Determine next steps: based on the results of the previous two tasks. A decision should be made to either finish the project and proceed to deployment or to initiate further iterations.
6. Deployment: application of the model in a production environment. In case a descriptive model is intended, application of its findings to the problem's objective.
    a. Plan deployment: summarize the deployment strategy, including not only the "what to do", but also the "how to do".
    b. Plan monitoring and maintenance: if the goal involves putting the plan into production, since models are not eternal and are data-dependent, monitorization and support for the model deployment should be planned.
    c. Final report: generate the final project report, and for accomplished projects, a comprehensive presentation of the project results.
    d. Review project: assess what went well and not so well to document the experience gained during the project.

An example of the use of CRISP-DM in tourism research can be seen in the paper written by Antonio et al. (2019) on how to use Big Data to predict cancellations in hotels.

2.4. Conclusion

We expect that reading this chapter may entice the reader not previously familiarized with Data Mining possibilities and methods to learn more about this topic and to apply it in e-Tourism research projects. We believe that data-driven research has the potential of enhancing E-Tourism research.

2.4.1. Challenges and limitations of Data Mining projects

Significant challenges and limitations on the application of Data Mining and Predictive Analytics research in e-Tourism still exist.

Let us begin by commenting on the access to the data. Although data is being generated at an unprecedented scale, much is privately-owned (e.g., hotels, airlines, and other company's data), which difficult the access to it. Even public data, such as user-generated content in social networks websites is now more challenging to use and subject to rules of use, like in the case of rules imposed by the European Union's General Data Protection Regulation (GDPR).

Another important point to be made regards to data quality is that data must be accurate, reliable, unbiased, timely, and appropriate for the problem under analysis.

Another key factor is a multi-disciplinary knowledge. Since Data Mining and Predictive Analytics combine techniques from multiple fields of study, including Statistics, Machine Learning, and Data Visualization, e-Tourism research teams, besides being composed of people with a background in social sciences, should also be composed of elements with a background in ICT or quantitative fields.

Lastly, one must consider the ethical implications. When personal or ethnical information is implicated, Data Mining and Predictive Analytics models could easily be used to discriminate. Profiling should carefully consider the possibility of results being used as a discriminatory tool, which is of the more significant concerns of GDPR. Even unintentionally, modeling could be incorrectly used. As an example, take a hotel website and a model to determine prices or cancellation policies dynamically. The use of zip code or country variables may be used to discriminate for or against who lives in certain regions typically inhabited by ethnicities, conditioning the diversity of the hotel customers. It is crucial that researchers, by design, always consider the ethical or privacy implications of their Data Mining and Predictive Analytics investigation.

2.4.2. Expected Future Developments

While Descriptive Analytics and Predictive Analytics are now commonly employed in business and increasingly employed for e-Tourism research, only a few examples are found in academic research of Prescriptive Analytics (Lepenioti et al. 2020). Prescriptive Analytics has the potential of enabling optimized decision making ahead of time, allowing users and modelers to comprehend the results of possible courses-of-actions or scenarios. Due to this potential, it is expected that, in future years, Prescriptive Analytics becomes one of the hottest topics within Data Mining research. For example, to create post-pandemic tourism scenarios or to simulate the impact on tourism of future crisis.

Cross-References

For more information on predictive modeling, see the chapter "Artificial intelligence and Machine Learning".

For more information on data visualization, see the chapter "Visual Methods and Visual Data Analysis in E-tourism Research".

For more information on big data and data sources available in e-Tourism the chapters "Big Data" and "Open and Commercial (big) Data in Tourism".

For more information about text mining, see the chapter "Natural Language Processing and Text Analytics in Tourism: Methodologies, Challenges and Opportunities".

References

Antonio N, Almeida A de, Nunes L, et al (2018) Hotel online reviews: different languages, different opinions. Information Technology & Tourism 18:157–185. https://doi.org/10.1007/s40558-018-0107-x

Antonio N, de Almeida A, Nunes L (2019) Big Data in hotel Revenue Management: Exploring cancellation drivers to gain insights Into booking cancellation behavior. Cornell Hospitality Quarterly 60:298–319. https://doi.org/10.1177/1938965519851466

Bach MP, Schatten M, Marušic Z (2013) Data mining applications in tourism: A keyword analysis. In: Hunjak, Tihomir, Lovrenčić, Sandra, Tomičić, Igor (eds) Proceedings of the 24th Central European Conference on Information and Intelligent Systems. Varaždin, Croatia, pp 26–32

Bermingham L, Lee I (2014) Spatio-temporal sequential pattern mining for tourism sciences. Procedia Computer Science 29:379–389. https://doi.org/10.1016/j.procs.2014.05.034

Brida JG, Disegna M, Osti L (2011) Residents' perceptions of tourism impacts and attitudes towards tourism policies in a small mountain community. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.1839244

Buhalis D, Law R (2008) Progress in information technology and tourism management: 20 years on and 10 years after the Internet—The state of eTourism research. Tourism Management 29:609–623. https://doi.org/10.1016/j.tourman.2008.01.005

Cai G, Hio C, Bermingham L, et al (2014) Mining frequent trajectory patterns and regions-of-interest from Flickr photos. In: 2014 47th Hawaii International Conference on System Sciences. pp 1454–1463

Chang K-C, Chen M-C, Kuo N-T, et al (2016) Applying data mining methods to tourist loyalty intentions in the international tourist hotel sector. Anatolia 27:271–274. https://doi.org/10.1080/13032917.2015.1099554

Chapman P, Clinton J, Kerber R, et al (2000) CRISP-DM 1.0: Step-by-step data mining guide. In: The Modeling Agency. https://the-modeling-agency.com/crisp-dm.pdf. Accessed 10 Sep 2015

Chen G, Bao J, Huang S (Sam) (2014) Segmenting chinese backpackers by travel motivations. International Journal of Tourism Research 16:355–367. https://doi.org/10.1002/jtr.1928

Chen Q, Hu Z, Su H, et al (2018) Understanding travel patterns of tourists from mobile phone data: A case study in Hainan. In: 2018 IEEE International Conference on Big Data and Smart Computing (BigComp). pp 45–51

Claveria O, Monte E, Torra S (2015) Tourism demand forecasting with neural network models: Different ways of treating information. International Journal of Tourism Research 17:492–500. https://doi.org/10.1002/jtr.2016

Delen D, Demirkan H (2013) Data, information and analytics as services. Decision Support Systems 55:359–363. https://doi.org/10.1016/j.dss.2012.05.044

Delen D, Sirakaya E (2006) Determining the efficacy of data-mining methods in predicting gaming ballot outcomes. Journal of Hospitality & Tourism Research 30:313–332. https://doi.org/10.1177/1096348006286795

Falk M, Vieru M (2018) Modelling the cancellation behaviour of hotel guests. International Journal of Contemporary Hospitality Management 30:3100–3116. https://doi.org/10.1108/IJCHM-08-2017-0509

Francalanci C, Hussain A (2016) Discovering social influencers with network visualization: evidence from the tourism domain. Information Technology & Tourism 16:103–125. https://doi.org/10.1007/s40558-015-0030-3

Guo Y, Barnes SJ, Jia Q (2017) Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. Tourism Management 59:467–483. https://doi.org/10.1016/j.tourman.2016.09.009

Halim MA, Saraf NM, Hashim NI, et al (2018) Discovering new tourist attractions through social media data: A case study in Sabah Malaysia. In: 2018 IEEE 8th International Conference on System Engineering and Technology (ICSET). IEEE, Bandung, Indonesia, pp 157–161

Han J, Kamber M, Pei J (2012) Data mining: Concepts and techniques, Third edition. Elsevier, Waltham, MA, USA

Hastie T, Tibshirani R, Friedman J (2017) The elements of statistical learning, Second Edition. Springer series in statistics Springer, Berlin

Höpken W, Eberle T, Fuchs M, Lexhagen M (2020) Improving Tourist Arrival Prediction: A Big Data and Artificial Neural Network Approach. Journal of Travel Research 0047287520921244. https://doi.org/10.1177/0047287520921244

Hsu L, Hsu C, Lin T (2010) Data mining in personalized travel information system. In: 2010 2nd International Conference on Information Technology Convergence and Services. pp 1–4

Hu F, Li Z, Yang C, Jiang Y (2019) A graph-based approach to detecting tourist movement patterns using social media data. Cartography and Geographic Information Science 46:368–382. https://doi.org/10.1080/15230406.2018.1496036

Kitchin R (2014) Big Data, new epistemologies and paradigm shifts. Big Data & Society 1:205395171452848. https://doi.org/10.1177/2053951714528481

Larose DT, Larose CD (2015) Data mining and predictive analytics, Second edition. John Wiley & Sons, Inc., Hoboken, N.J.

Law R, Mok H, Goh C (2007) Data mining in tourism demand analysis: A retrospective analysis. In: Alhajj R, Gao H, Li J, et al. (eds) Advanced Data Mining and Applications. Springer Berlin Heidelberg, pp 508–515

Lepenioti K, Bousdekis A, Apostolou D, Mentzas G (2020) Prescriptive analytics: Literature review and research challenges. International Journal of Information Management 50:57–70. https://doi.org/10.1016/j.ijinfomgt.2019.04.003

Li G, Law R, Vu HQ, et al (2015) Identifying emerging hotel preferences using Emerging Pattern Mining technique. Tourism Management 46:311–321. https://doi.org/10.1016/j.tourman.2014.06.015

Li S, Hao J, Chen Z (2010) Graph-based service quality evaluation through mining web reviews. In: Proceedings of the 2010 International Conference on Natural Language Processing and Knowledge Engineering. IEEE, Beijing, China, pp 280–287

Liao S, Chen Y-J, Deng M (2010) Mining customer knowledge for tourism new product development and customer relationship management. Expert Systems with Applications 37:4212–4223. https://doi.org/10.1016/j.eswa.2009.11.081

Maimon O, Rokach L (eds) (2010) Data Mining and Knowledge Discovery handbook, Second edition. Springer, Boston, MA

Malik MM, Abdallah S, Ala'raj M (2018) Data mining and predictive analytics applications for the delivery of healthcare services: a systematic literature review. Ann Oper Res 270:287–312. https://doi.org/10.1007/s10479-016-2393-z

Mariani M, Baggio R, Fuchs M, Höepken W (2018) Business intelligence and big data in hospitality and tourism: a systematic literature review. International Journal of Contemporary Hospitality Management 30:3514–3554. https://doi.org/10.1108/IJCHM-07-2017-0461

Mazzocchi F (2015) Could Big Data be the end of theory in science? EMBO reports 16:1250–1255. https://doi.org/10.15252/embr.201541001

Moro S, Rita P (2016) Forecasting tomorrow's tourist. Worldwide Hospitality and Tourism Themes; Bingley 8:643–653

Muresan I, Oroian C, Harun R, et al (2016) Local residents' attitude toward sustainable rural tourism development. Sustainability 8:100. https://doi.org/10.3390/su8010100

Pei S (2013) Application of data mining technology in the tourism product's marketing CRM. In: 2013 2nd International Symposium on Instrumentation and Measurement, Sensor Network and Automation (IMSNA). IEEE, Toronto, Canada, pp 400–403

Pesonen J, Laukkanen T, Komppula R (2011) Benefit segmentation of potential wellbeing tourists. Journal of Vacation Marketing 17:303–314. https://doi.org/10.1177/1356766711423322

Rossetti M, Stella F, Zanker M (2016) Analyzing user reviews in tourism with topic models. Inf Technol Tourism 16:5–21. https://doi.org/10.1007/s40558-015-0035-y

Sánchez-Martín J-M, Rengifo-Gallego J-I, Blas-Morato R (2019) Hot spot analysis versus cluster and outlier analysis: An enquiry into the grouping of rural accommodation in Extremadura (Spain). ISPRS International Journal of Geo-Information 8:176. https://doi.org/10.3390/ijgi8040176

Shapoval V, Wang MC, Hara T, Shioya H (2017) Data mining in tourism data analysis: Inbound visitors to japan. Journal of Travel Research 0047287517696960. https://doi.org/10.1177/0047287517696960

Shmueli G (2010) To explain or to predict? Statistical Science 25:289–310. https://doi.org/10.1214/10-STS330

Srivihok A, Intrapairot A (2016) To be or not be competitive country: Analysis of travel and tourism competitiveness index by multiple data mining techniques. In: 2016 6th International Workshop on Computer Science and Engineering, WCSE 2016. International Workshop on Computer Science and Engineering (WCSE), Tokyo, Japan, pp 206–213

Strasser BJ (2012) Data-driven sciences: From wonder cabinets to electronic databases. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 43:85–87. https://doi.org/10.1016/j.shpsc.2011.10.009

Witten IH, Frank E, Hall MA (2011) Data mining: practical machine learning tools and techniques, 3rd ed. Morgan Kaufmann, Burlington, MA

Wu EHC, Law R, Jiang B (2010) Data mining for hotel occupancy rate: An Independent component analysis approach. Journal of Travel & Tourism Marketing 27:426–438. https://doi.org/10.1080/10548408.2010.481585

Xie H, Tang W (2009) Application research of Data Mining in travel agency's Customer Relationship Management. In: Li Q, Yu F, Liu Y, Russell M (eds) 2009 Second International Workshop on Computer Science and Engineering. IEEE, Qingdao, China, pp 464–467

Zaki MJ, Meira W (2014) Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press, New York, NY, USA

Zhao X, Ji K (2013) Tourism e-commerce recommender system based on web data mining. In: 2013 8th International Conference on Computer Science Education. pp 1485–1488