Decision Support Based on Data Mining for Post COVID-19 Tourism Industry

Nenad Petrovic
University of Niš

Submitted for XV International SAUM Conference
on Systems, Automatic Control and Measurements
Niš, Serbia, September 9th-10th, 2021

Decision Support Based on Data Mining for Post COVID-19 Tourism Industry

Nenad Petrović, Vasja Roblek and Nino Papachashvili

[+] Department of Computer Science, University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 14, 18000 Niš,
E-Mail: nenad.petrovic@elfak.ni.ac.rs

[++] Faculty of Organisation Studies, Ulica talcev 3, 8000 Novo mesto, Slovenia, E-mail: vasja.roblek@gmx.com

[+++] Sulkhan-Saba Orbeliani University, Institute for Development Studies, Tbilisi, Georgia. E-mail: n.papachashvili@sabauni.edu.ge

Abstract* – The ongoing coronavirus pandemic has huge impact on almost every industry sector. However, its consequences caused disastrous losses when it comes to tourism. Strict government measures, such as lockdown, public gathering and travel prohibition, together with temporary closure of hospitality objects have not only reduce the number of tourists dramatically, but also led to huge number of jobs lost in tourism and hospitality. In this paper, we examine how the commonly used data mining techniques can contribute to revival of tourism industry under current circumstances. As outcome, the implementation of four case studies using RapidMiner tool is presented: 1) reservation cancellation detection based on classification 2) tourist number prediction with respect to number of COVID-19 cases leveraging regression 3) Cross-selling of additional services relying on association rule mining 4) segmentation of tourists regarding their spending and services using clustering.

***Key words:* coronavirus, data mining, RapidMiner, tourism**

I. INTRODUCTION

During the third industrial revolution in the 1960s, humanity underwent a social transformation that led to the emergence of the information society. It was a new form of social existence in which the primary task was to collect, store, analyse, and share networked information [1]. Technological development during Industry 4.0 enabled the transformation from a service-oriented society to a human-centred technology, and with IoT and Big Data, various industries and the human social environment have entered the process of informatisation. Informatisation has created the cyber-physical environment (CPE) and Big Data, enabling the information society to connect intangible assets as information networks. Industry 4.0 represents a whole new way of integrating technology into society. New technologies are being researched and developed that connect the physical, digital and biological worlds. These new technologies affect all disciplines, economies and industries [3]. Industry 4.0 appears as a continuation of the third industrial revolution. It enables the digital interconnection of products, machines, tools and more. It brings 3D printers, self-driving vehicles, AI, and nanotechnology, but unlike the second and third industrial revolutions, which were based on raw materials and energy, Industry 4.0 emphasises knowledge as an essential resource [4].

With the digitisation of our society, the amount of data collected and stored in many data centres worldwide is increasing exponentially. Since the amount of data collected has long exceeded the human ability to discover useful information, new techniques and tools had to be developed to discover new knowledge from this data. In this article, we focus on data mining in the tourism industry.

In the last thirty years, the digitalisation of the tourism industry has brought revolutionary changes in the technological development of the business of tourism companies, which are increasingly becoming an added value of tourism [5]. In the 21st century, all communication between the members of the tourism value chain (travel agencies, hotels, hospitality industry, airlines, railway companies, etc.) and tourists has moved to the Internet. The development of the tourism industry is inextricably linked to the processing of tourist information within the framework of digital solutions, which have become a tool for strengthening tourism competition. The growth of data has also led to creating a huge information space in the tourism industry. It has become a guide to the use of Big Data in tourism, showing how to fully exploit the vast original of tourism data and mine and analyse big tourism data reflecting information quickly, accurately and conveniently [6].

The concept of Data Mining, which refers to knowledge discovery and databases, first appeared in the late 1980s. Han, Kamber and Pei define data mining as a process of discovering interesting patterns and knowledge based on the analysis of a large amount of data available in various repositories [7]. Data mining includes linkage analysis, sequence pattern, classification, grouping, irregularity detection etc. For example, when using tourism data, aggregation analysis is used to search for tourism data and look for patterns with a high probability of occurrence or analyse the similarity of tourism data by aggregating and sorting data and storing similar data together [8].

Decision-makers, the tourism industry and government administrations, therefore, need data on the relationship between tourism activities and tourism preferences for their work. These data are used to support planning for tourism infrastructure development. In the context of operational and strategic decisions, managers need analysis on staffing levels, investments, aircraft utilisation, hotel service utilisation, etc. [9]. Data mining based on machine learning has replaced the classical statistical methods in the tourism industry, where the assumptions about the distribution of data must first be established [10]. The main problem of statistical analysis is that in case of violation of the assumptions, the validity of the data cannot be guaranteed [11]. In practice, data mining in the tourism industry is most often used for predicting tourist spending, analysis of tourist profiles, and prediction of the number of tourists [12].

The recent COVID-19 pandemic has caused enormous losses in tourism sector, especially due to travel limitations [13]. This paper explores how data mining techniques can be leveraged in order to support decision-making regarding the post COVID-19 tourism and its recovery under current circumstances. As outcome, we introduce four case studies in context of tourism industry implemented using RapidMiner tool.

II. BACKGROUND

A. Decision Support Based on Data Mining Techniques

In Fig. 1, a typical workflow leveraging data mining techniques for decision support is displayed.

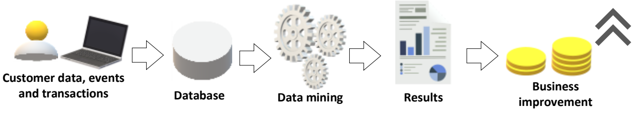


Fig. 1. Workflow of decision support based on data mining in business information systems.

The data about customers, transactions and events that occur is collected and stored into database. Later, the previously acquired data is analyzed using data mining techniques in order to extract knowledge and hidden patterns. Moreover, the results are interpreted in particular business context, aiming to achieve the improvement regarding some of the relevant aspects, such as profit increase, loss reduction or customer number increase. In Table I, an overview of commonly used data mining techniques and their typical use cases across various domains is given [15-18].

TABLE I

Commonly used data mining techniques and their use cases

Name	Description	Use cases
Classification	Predicting the class where the observed sample belongs to	Churn prediction - Detecting customers likely to cancel a subscription to a service
Regression analysis	Modelling the relationship between scalar response (output, dependent variable) and one or many inputs (independent variables)	Numerical (real-valued) outcome
		Cryptocurrencies prices
		Stock prices
		Energy consumption
		Number of COVID-19 cases
Association rule mining (market basket analysis)	A process that looks for relationships of objects that "go together" within same business context	Product placement
		Physical shelf arrangement
		Cross-selling
		Customer retention
Clustering	Dividing the dataset into a number of groups, so the samples from the same group are more similar to others in the same group, while they are dissimilar to samples from other groups	Market segmentation
		Fraud detection

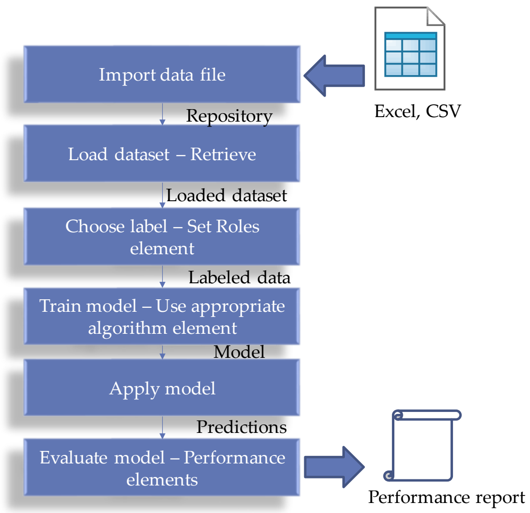
Use cases for these techniques are depicted in Fig. 2.

Fig. 2. Data mining case study illustration.

B. RapidMiner

RapidMiner [14] is a data science software platform that provides an integrated environment for data preparation, analysis and visualization, initially released in 2006. Despite the fact that paid license is required for full features and support, free edition with certain limitations (10 000 rows, 1 logical processor) also exists, which makes it accessible for wider audience. It covers both data mining algorithms and different types of neural networks. Furthermore, RapidMiner offers intuitive visual interface that does not require any coding, which makes it suitable even for domain experts working in tourism sector. The components are available as blocks (also called *operators* in RapidMiner Studio) that are connected in order to make data analysis process. Each of these block contains a set of adjustable parameters that are relevant to the algorithm they refer to. In Fig. 3, the illustration of data analysis workflow in RapidMiner is shown.

Fig. 3. RapidMiner data analysis workflow.



III. CASE STUDIES

In this section, four case studies adopting different data mining techniques in tourism industry are presented. The datasets used in case studies were constructed based on survey carried out in region of Pirot and Stara Planina, Serbia covering the period of January 2020 – April 2021. RapidMiner process flows are available within the author's GitHub repository[1].

[1] https://github.com/penenadpi/tourism_rapidminer

A. Classification - Reservation Cancellation Prediction

This case study is based on binary classification, inspired by churn prediction approach. Its goal is to answer whether the customer is likely to cancel travel reservation. After identification of such customers, additional discounts or free services (such as free COVID-19 test if necessary for their target destination or free excursion) can be offered in order to motivate them and make more confident about their travel arrangement, preventing additional losses to agencies, hotels and other parties. Several factors are taken as input: 1) month – which illustrates the travelling season, as some destinations might be more popular at certain part of the year, which reduces the possibility of cancellation 2) completed reservations – this number reflects how often the considered customer travels 3) cancelled reservations – factor that illustrates how many times the customer decided to cancel trip reservation in past 4) COVID-19 cases – current number of active cases in target country. The output is binary value which takes value “1” if reservation is going to be cancelled, while it is “0” otherwise. In Table II, the header of dataset used in this case study is given.

TABLE II

Reservation cancellation prediction dataset structure

Month	Completed reservations	Cancelled reservations	COVID-19 cases	Cancel [0/1]

When it comes to implementation in RapidMiner (process flow given in Fig. 4), k-Nearest Neighbours (*k-NN*) [19] classification algorithm was used.

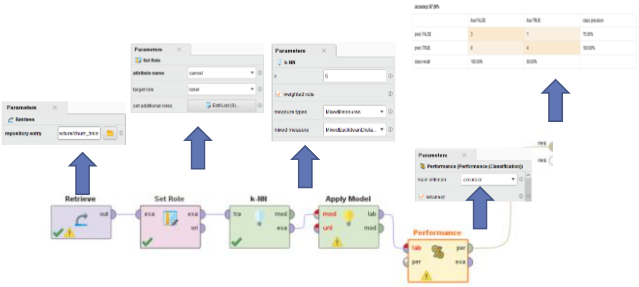


Fig. 4. RapidMiner process flow for classification using k-NN.

This algorithm determines the class of new observation with based on majority class of its *k* closest samples (which was 5 in our case), as illustrated in Fig. 5. The distance between samples was calculated using Euclidean distance.

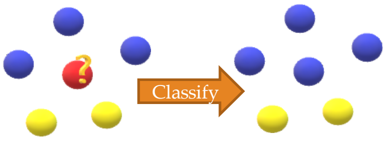


Fig. 5. k-NN classification algorithm illustration for k=5.

When it comes to prediction performance, the achieved accuracy (number of correctly predicted samples against total number of observations) was 87.5%.

B. Regression – Tourist Number Prediction

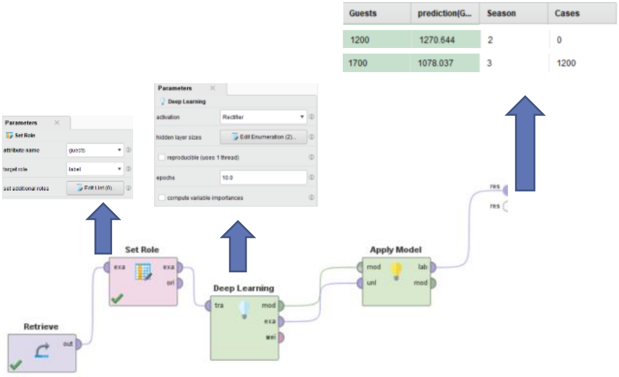
The aim of this case study is to predict the number of tourists within certain region considering the active COVID-19 cases and current season using regression approach. Table III shows the header of dataset used in this case study.

TABLE III

Tourist number prediction dataset structure

Season	COVID-19 cases	Guests [num]

For this case study, a deep learning neural network [20] was adopted, as other methods in RapidMiner not able to handle the non-linearities properly, leading to poor performance. It has two hidden layers with 50 nodes and ReLU activation function, while Adam optimizer was used. The training was done in 10 epochs. In Fig. 6, RapidMiner flow for regression based on deep learning.



Guests	prediction(G...	Season	Cases
1200	1270.644	2	0
1700	1078.037	3	1200

Parameters
Set Role
attribute name: guests
target role: label
set additional roles: Edit List (0)

Parameters
Deep Learning
activation: Rectifier
hidden layer sizes: Edit Enumeration (2)...
☐ reproducible (uses 1 thread)
epochs: 10.0
☐ compute variable importances

Fig. 6. RapidMiner process flow for deep learning-based regression.

Regarding the prediction performance, relative error of this model ranged from 38% to 4% in our experiments.

C. Association Rule Mining – Cross-Selling in Tourism

On the other side, the purpose of this case study is to identify the additional services, products of local craftmanship and excursions that are often bought together by tourists visiting the same destination. According to this information, the marketing department of tourist agency can construct more attractive travel arrangement bundles, while supporting local craftmanship at the same time. In this context, we make use of apriori [19] algorithm for association rule mining. The structure of dataset is shown in Table IV.

TABLE IV

Excursion cross-selling dataset structure

Invoice id	Service 1	Season	Amount

In Fig.7, the process flow behind this case study is shown.

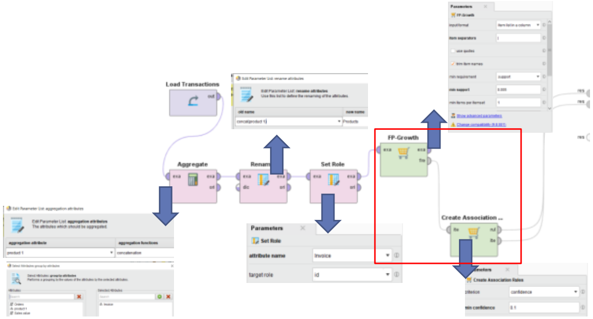


Fig. 7. RapidMiner process flow for rule mining using apriori.

Before the algorithm execution, aggregation of data is performed, so all the services are appended to the same tourist invoice. Moreover, the frequent item set is generated with 0.005 value of minimum support. Finally, the set of association rules is discovered with value 0.1 of minimum confidence parameter. In our case study, for example, it was discovered that foreign tourists in Pirot region frequently visit Museum of Ponišavlje, while buying local specialties - cheese and ironed sausage.

D. Clustering – Tourist Segmentation

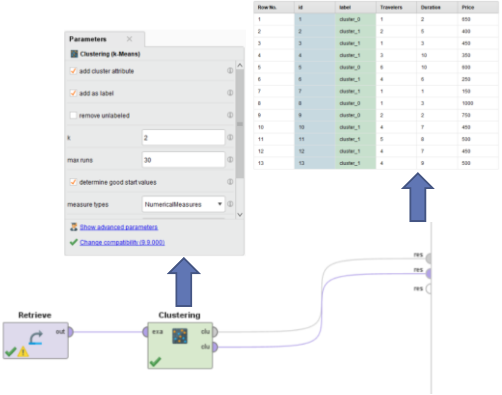
Finally, the goal of this case study is to provide mechanism which will recognize different types of tourists, which can help the agency regarding the offers of additional services and excursions. For this purpose, we adopt k-means clustering algorithm [19]. The following features are considered (Table V): 1) Travelers – number of people within the travel reservation 2) Duration – how many days the reservation lasts 3) Total – the overall price of the reservation. In this case, output represent the assigned number of corresponding cluster. We set the desired number of clusters as 2.

TABLE V

Tourist segmentation dataset structure

Travelers	Duration	Total	Cluster [0/1]

In Fig. 8, a screenshot of RapidMiner clustering process flow is shown.



Row No.	id	label	Travelers	Duration	Price
1	1	cluster_0	1	2	650
2	2	cluster_1	2	5	400
3	3	cluster_1	1	3	450
4	4	cluster_1	3	10	350
5	5	cluster_0	6	10	600
6	6	cluster_1	4	6	250
7	7	cluster_1	1	1	150
8	8	cluster_0	1	3	1000
9	9	cluster_0	2	2	750
10	10	cluster_1	4	7	450
11	11	cluster_1	5	8	500
12	12	cluster_1	4	7	450
13	13	cluster_1	4	9	500

Parameters

Clustering (k-Means)

☑ add cluster attribute

☑ add as label

□ remove unlabeled

k: 2

max runs: 30

☑ determine good start values

measure types: NumericalMeasures

Show advanced parameters

Change compatibility (9.9.000)

Fig. 8. RapidMiner process flow for clustering using k-means.

Considering the obtained output, we can interpret the clusters as following: 0) business travels 1) family/couple travels. When it comes to family/couple arrangements, the number of people is at least two, while duration is longer and total spending lower. On the other side, for business travels, the number of persons is usually up to 2, stay duration is shorter, but involving higher costs.

IV. CONCLUSION AND FUTURE WORK

According to the achieved results, it can be concluded that data mining techniques show enormous potential when it comes to decision support in context of post COVID-19. They not only improve the existing business processes in tourism sector, but enable novel services as well. In future, we plan to adopt this approach in order to support the development of tourism in Georgia. Moreover, data mining knowledge transfer sessions (workshops, tutorials) for students in hospitality and tourism industry are scheduled. Finally, we also consider the adoption of data mining within augmented reality mobile apps targeting tourists [21].

ACKNOWLEDGMENT

This work has been supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia.

REFERENCES

[1] F. Webster, *Theories of the information society,* 4th ed., London, Routledge, 2014.

[2] V. Roblek, M. Meško, M. P. Bach, O. Thorpe & P. Šprajc, “The interaction between internet, sustainable development, and emergence of society 5.0”, Data, 5(3), 80, 2020.

[3] L. Caruso, “Digital innovation and the fourth industrial revolution: epochal social changes?” Ai & Society, 33(3), pp. 379-392, 2018.

[4] B. Trenovski, & G. Merdzan, “Lessons learned from the fourth industrial revolution for the global economy”, Knowledge International Journal, 43(1), pp. 89-95, 2020.

[5] J. Navío-Marco, L. M. Ruiz-Gómez, C. Sevilla-Sevilla, “Progress in information technology and tourism management: 30 years on and 20 years after the internet-Revisiting Buhalis & Law's landmark study about eTourism”, Tourism management, 69, pp. 460-470, 2018.

[6] P. Centobelli, V. Ndou, “Managing customer knowledge through the use of big data analytics in tourism research”, Current Issues in Tourism, 22(15), pp. 1862-1882, 2019.

[7] J. Han, M. Kamber, J. Pei, “Data mining concepts and techniques”, third edition, The Morgan Kaufmann Series in Data Management Systems, 5(4), pp. 83-124, 2011.

[8] K. Zhang, Y. Chen, C. Li, “Discovering the tourists' behaviors and perceptions in a tourism destination by analysing photos' visual content with a computer deep learning model: The case of Beijing”, Tourism Management, 75, 595-608, 2019.

[9] G. Xie, Y. Qian, S. Wang, “Forecasting Chinese cruise tourism demand with big data: An optimised machine learning approach”, Tourism Management, 82, 104208, 2021.

[10] H. Huang et al., “Location based services: ongoing evolution and research agenda”, Journal of Location Based Services, 12(2), pp. 63-93, 2018.

[11] M. T. Leung, S. Pan and M. Sun, “A review of data analytic methods and their applications in e-commerce research”, *Journal of Current Issues in Media & Telecommunications,* 9(2-3), pp. 117-176, 2017.

[12] S. J. Miah, H. Q. Vu, J. Gammack, M. McGrath, “A big data analytics method for tourist behaviour analysis”, Information & Management, 54(6), pp. 771-785, 2017.

[13] Tourism Suffered Massive Losses In 2020 [online]. Available on: https://www.statista.com/chart/24681/travel-and-tourisms-contribution-to-gdp/, last accessed: 21/05/2021.

[14] RapidMiner [online]. Available on: https://rapidminer.com/get-started/ , last accessed: 21/05/2021.

[15] N. Petrović, “Adopting Data Mining Techniques in Telecommunications Industry: Call Center Case Study”, IEEESTEC – 11th Student Projects Conference, pp. 11-14, 2018.

[16] N. Petrovic, “Churn Prediction in Telco Industry Leveraging Call Center Data”, IcETRAN 2019, 845-850, 2019.

[17] N. Petrović, Đ. Kocić, “Data-driven Framework for Energy-Efficient Smart Cities”, Serbian Journal of Electrical Engineering, Vol. 17, No. 1, Feb. 2020, pp. 41-63, 2020.

[18] N. Petrović, “Simulation Environment for Optimal Resource Planning During COVID-19 Crisis”, ICEST 2020, pp. 23-26, 2020.

[19] J. Han, M. Kamber, J. Pei, *Data Mining: Concepts and Techniques*, third edition, Morgan Kaufmann Publishers, 2012.

[20] Y. Bengio, “Learning Deep Architectures for AI”, Foundations and Trends in Machine Learning, Vol. 2, No. 1, pp. 1-127, 2009.

[21] N. Petrović, V. Roblek, M. Khokhobaia, I. Gagnidze, “AR-Enabled Mobile Apps to Support Post COVID-19 Tourism”, unpublished, TELSIKS 2021.