# Introduction

## General Overview

### Global Perspective

The travel and tourism industry plays a significant role in the global economy, generating 10.4% of the world GDP and supporting over 319 million jobs as reported by the World Economic Forum in 2018 [21, 22, 23]. Many international travelers rely on digital means for planning their trips and base their decisions on the shared online information from other travelers. To improve customer satisfaction and increase repeat business, it is essential to meet their expectations. Big Data analytics has been crucial in achieving this goal by utilizing the vast amount of Online User-Generated Content (UGC) as a valuable resource for companies and governments to shape their tourism strategies [24, 25, 26].

The use of algorithms from the Artificial Intelligence field, specifically from the Natural Language Processing (NLP) area, is crucial for handling Online User-Generated Content (UGC). NLP is a sub-field of AI that focuses on analyzing and representing natural language text, with the goal of achieving human-like processing capabilities. This area of AI has its roots in the intersection of artificial intelligence and linguistics, and has grown in interest since its beginnings in 1950 [27]. NLP techniques are applied to various levels of linguistic examination for different purposes [28].

### National Perspective

Natural language processing (NLP) techniques can play a significant role in the tourism industry by improving customer service through chatbots, virtual assistants and sentiment analysis [29, 30], increasing efficiency by automating the process of extracting information from unstructured data, such as emails, social media posts, and customer reviews and personalizing tourism-related recommendations for individuals based on their preferences and past experiences. NLP can also generate summaries of customer reviews, making it easier for businesses to quickly identify patterns and trends in customer feedback. In some countries, government may provide support for the development and use of NLP in the tourism industry through funding and research grants, and existing partnerships between academia and industry in the field of NLP and tourism may exist, where researchers work closely with tourism businesses to develop and implement NLP-based solutions. However, challenges such as lack of data infrastructure, lack of knowledge or skills in NLP among tourism businesses may be faced by countries.

## Research Background

## Problem Statement

Tourism is an important industry that generates significant revenue for many countries. However, travelers often face a challenge in finding the best tourist areas due to the large amount of information available online. The current process of finding information about tourist areas often involves manually searching through multiple websites and reading through various tourism reports written in natural language, which can be a tedious and time-consuming task. The unstructured nature of the information makes it difficult for travelers to compare different tourist areas based on specific criteria such as location, cost, and activities, leading to a lack of confidence in the final decision and ultimately affecting the travelers' experience. There is a need for a solution that can automatically process large amounts of unstructured data, such as tourism reports, and extract relevant information about different tourist areas, this information can then be used to classify or group tourist areas based on their features or overall sentiment, making it easier for travelers to find the best tourist areas for their needs, and also potentially improve the overall tourism industry by providing more accurate and relevant information to travelers.

## 1.4 Research aim and objectives

The main aim of this research is to develop a solution that uses natural language processing (NLP) techniques to help travelers find the best tourist areas based on tourism reports. The specific objectives of this research are as follows:

* To collect and clean a large dataset of tourism reports written in natural language
* To use NLP techniques such as tokenization, stemming, and sentiment analysis to extract relevant information from the reports
* To use machine learning algorithms such as clustering or classification to group the tourist areas based on their features or to classify them based on their overall sentiment
* To evaluate the performance of the model using the collected dataset

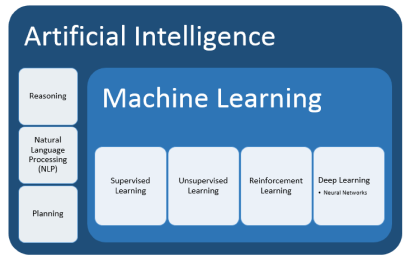
The expected outcome of this research is a solution that can process large amounts of unstructured data, such as tourism reports, and extract relevant information about different tourist areas. By providing accurate and relevant information, this solution aims to help travelers plan their trips more efficiently and find the best tourist areas for their needs. Additionally, the solution could also potentially improve the overall tourism industry by providing more accurate and relevant information to travelers, leading to better-informed decisions, and ultimately, more satisfactory traveling experiences.

# chapter 2

## Natural Language Processing

### Definition and Distinction

In this section, we will discuss the relationship between Artificial Intelligence (AI), Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL) and how they all fall under the broader field of AI. AI is a broad field of study that aims to create systems that can "think" and perform tasks that would normally require human intelligence, such as understanding natural language, recognizing images, and making decisions. NLP is a subfield of AI that focuses on the interaction between computers and human language. It involves the use of algorithms and statistical models to analyze, understand, and generate human language [3]. NLP tasks include text classification, language translation, sentiment analysis, speech recognition, and more. ML is a method of teaching computers to learn from data, without being explicitly programmed. It is a key component of AI and is used in various NLP tasks. DL is a subset of ML that uses neural networks with multiple layers to analyze and understand data. DL has been successful in many NLP tasks, such as language translation, text summarization, and sentiment analysis.



[1]

### Techniques and methods

NLP techniques and methods can be broadly categorized into the following areas:

**Tokenization**

Tokenization is the process of breaking down text into individual words (tokens), phrases, or sentences. Tokens are the smallest independent units of natural language, typically representing individual words [2]. It is the first step in many NLP tasks, such as text classification, sentiment analysis, and language translation. There are different techniques for tokenization, such as using regular expressions, using rule-based algorithms, or using machine learning algorithms. A simple tokenizer divides the sentence "Natural Language Processing is a complex and interdisciplinary field" into the following nine tokens: "Natural", "Language", "Processing", "is", "a", "complex", "and", "interdisciplinary", "field".

**Stemming and Lemmatization**

Stemming and lemmatization are techniques used to reduce words to their base form. The goal is to reduce words to their root or dictionary form in order to reduce the dimensionality of the text and improve the performance of NLP models. Stemming is a crude heuristic process that chops off the end of a word [4], while Lemmatization is a more sophisticated process that uses morphological analysis to determine the base form of a word [5].

**Part-of-speech tagging**

Part-of-speech tagging is the process of identifying the grammatical role of each word in a sentence. It is used in many NLP tasks such as parsing, named entity recognition, and text summarization. There are different techniques for part-of-speech tagging, such as rule-based algorithms [6], statistical methods [7], and machine learning algorithms [8].

**Parsing**

Parsing is the process of analyzing the syntactic structure of a sentence. It is used to extract grammatical structure from text, which is important for many NLP tasks such as named entity recognition, text summarization, and machine translation. There are different techniques for parsing, such as rule-based algorithms [9], statistical methods [10], and machine learning algorithms [11], and recent neural network based models [12].

**Named entity recognition**

Named entity recognition is the process of identifying and classifying named entities such as people, organizations, and locations in text. It is used in many NLP tasks such as text summarization, question answering, and text-to-speech synthesis. There are different techniques for named entity recognition, such as rule-based algorithms [13], statistical methods [14], and machine learning algorithm [15] and recent neural network based models [16].

**Sentiment analysis**

Sentiment analysis is the process of determining the emotional tone of text. It is used in many NLP tasks such as opinion mining, reputation management, and text summarization. There are different techniques for sentiment analysis, such as lexicon-based methods, rule-based methods, and machine learning algorithms.

**Text summarization**

Text summarization is the process of creating a shorter version of a text that retains its main ideas. It is used in many NLP tasks such as text classification, sentiment analysis, and question answering. There are different techniques for text summarization, such as extractive summarization and abstractive summarization.

**Machine Translation**

Machine Translation is the process of translating text from one language to another. It is used in many NLP tasks such as text classification, sentiment analysis, and question answering. There are different techniques for Machine Translation, such as rule-based methods, statistical methods and neural machine translation.

**Word Embedding**

Word embedding is the process of representing words as high-dimensional vectors that capture their semantic meaning. It is used in many NLP tasks such as text classification, sentiment analysis, and question answering. There are different techniques for word embedding, such as Count-based methods, Predictive methods and Hybrid methods.

**Language Modelling**

Language Modelling is the process of estimating the probability distribution of sequences of words, which is used in many NLP tasks such as speech recognition, machine translation and text generation. There are different techniques for Language Modelling, such as n-gram model, recurrent neural network and

## Machine Learning

## Types of machine learning

The study of ML encompasses a wide range of topics and draws inspiration from other domains, including AI. Different types of ML as shown in Figure 3. Due to its complexity, machine learning has been separated into two main categories: supervised learning and unsupervised learning, and two ancillary categories: semi-supervised learning and reinforcement learning [28]. Each one has a distinct goal and course of action that produces outcomes and uses different types of data.

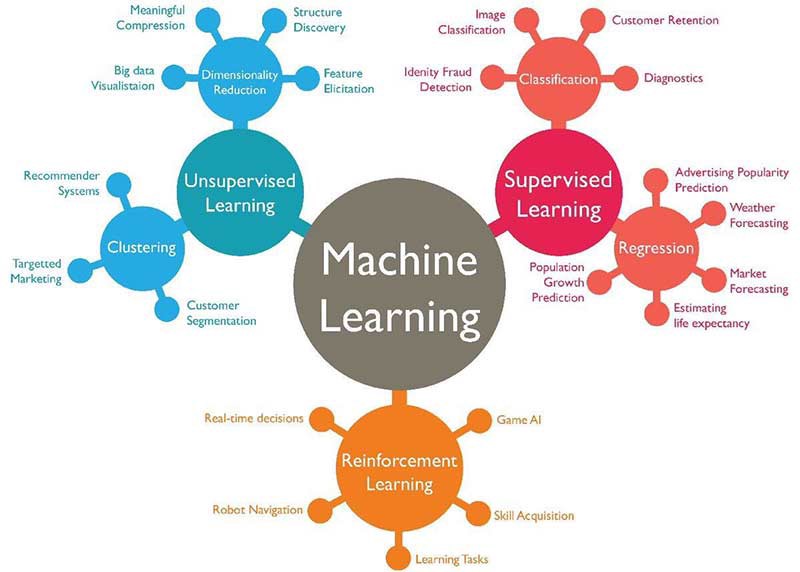


Figure 3 : Types of machine learning algorithms.

### Supervised learning

In supervised learning, we know the correct output and the relationship between input and output in this phase. It consists of labeled or annotated datasets and algorithms [46]. The model tries to find the best mapping function Y = F(X) for X and Y are input and output data respectively, Y can be a categorical or continuous value. The model generates a prediction for new unknown data. In supervised learning, the supervisor instructs the learning system on the labels or classes to associate with the labeled data [46]. Supervised learning algorithms induce models from these training data and these models can be used to classify the unlabelled test data [45]. Classification approaches such as KNN, SVM, Naive Bayes, Decision tree, and Random forest, are all examples of supervised learning [47].

### Unsupervised learning

Unsupervised learning is the opposite of supervised machine learning. In this case, we have an unlabeled dataset: we know the input data, but neither the output data nor the mapping function is known [46]. In this model, the machine observes the algorithms and finds the structure of the data. It has less computational complexity and uses real-time analysis of data through this model. Some frequently unsupervised learning algorithms are K-means, PCA, etc [47].

### Semi-Supervised learning

Semi-supervised learning also known as hybrid learning lies between supervised and unsupervised learning. Labeled data are sometimes time-consuming, and expensive and need annotation experts. However, unlabeled data are easy to collect [49]. This model uses a large amount of labeled and unlabeled data [48] to build accurate classifiers because it requires less human effort and gives higher accuracy. The goal of this semi-supervised learning is to understand how this combination can improve learning behavior [48]. The procedure is that the algorithm first uses unsupervised learning algorithms to cluster the unlabeled data and then uses the supervised learning algorithm [48]. These algorithms do better when dealing with large amounts of unlabelled data and fewer labeled data [47].

## Machine Learning Algorithms

### Decision Tree Algorithm

A Decision Tree (DT) is a supervised learning technique used to solve classification or regression problems [42]. It is a graphical representation similar to a tree. It starts with the root node, which expands on further branches and constructs a tree-like structure [42]. The features of the dataset are present by the internal nodes [42], the decision rules presented by the branches, and each leaf node (terminal node) represents the outcome (categorical or continuous value) of these decisions and do not contain other branches like decision nodes [41]. A DT simply asks a question, and based on the answer (Yes/No), it further split the tree into sub-trees. Figure 4 illustrates the DT structure.

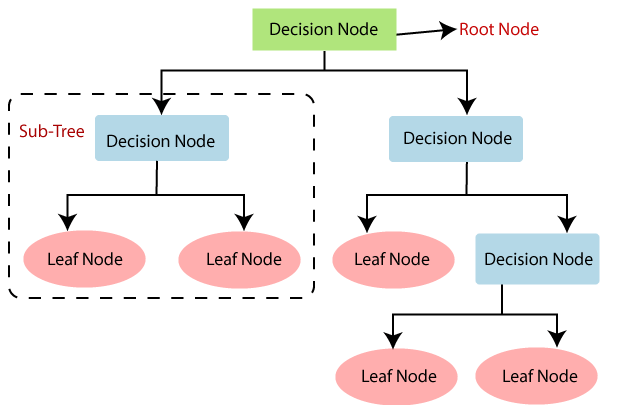


Figure 4: Decision Tree structure.

### Random Forests Algorithm

Random Forest [40] is a supervised ML algorithm that can be used for classification or regression problems. It is an ensemble learning method that combines some DT classifiers on random samples of a given dataset to solve complex problems and improve the prediction performance of the model [43]. It works in two stages, it first creates the random forest by combining some DTs then, it makes predictions for each decision tree. Based on a random selection of data samples, these algorithms create decision trees and obtain predictions from each tree. They then vote to determine which viable option is the best. It takes the average of these DTs in case of regression and their majority vote in case of classification. The below diagram (Figure 5) explains the working of the Random Forest algorithm.

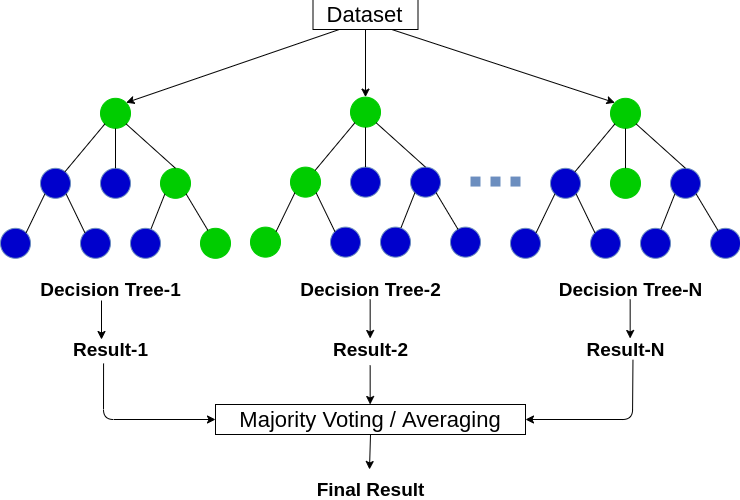


Figure 5: Random Forest algorithme.

### K-Nearest Neighbours

K-Nearest Neighbours (KNN) is a supervised ML algorithm used for classification and regression problems. It is a non-parametric algorithm that does not make any assumptions about underlying data. It is called also a lazy learner algorithm that it does not learn from the training set immediately instead of storing the dataset and at the classification phase, it acts on the dataset. Firstly, it calculates the Euclidean distance of K neighbors, among these k neighbors, it counts the number of data points in each category. Lastly, it classifies the new data into the most similar category. Figure 6 simulates the dataset distribution before and after using the KNN algorithm.

## Literature Review

Over the past decade, numerous research articles on tourism solutions have been published. One strategy to summarize the information of these research articles is by utilizing a Systematic Review (SR). SR is a powerful method to structure knowledge [31, 32]. For example, [33] reviewed 80 articles on the sports-tourism topic to identify the state-of-the-art in the field. The methodology employed for the review was developed by the author and is precisely explained to allow replicability.

Similarly, [34] investigated the link between tourism and the introduction of non-native species in marine environments, revising 32 articles. They employed the Collaborative Environmental Evidence methodology. [35] examined the state of research on innovation in tourism by reviewing 152 papers. They performed a bibliometric analysis to show the theoretical foundations of the studies. With clusters of the co-citation in the documents, they identified some trends that characterize the field. They did not mention the review methodology employed. [36] reviewed 144 articles on using the Big Data paradigm in tourism research. They wanted to understand the full-scale types of big data, data characteristics, analytic techniques, and touristic issues addressed in this field. The methodology followed for the review was not specified, but the implemented procedure is well explained for replicability. [37] studied the collaborative innovation in tourism with a review of 79 articles. They considered the studies’ location, the perspective of analysis, the employed methodology, the level of research, and the specific themes addressed. The reviewing methodology’s name was not specified, but the process was thoroughly explained. [38] scrutinized 17 articles to understand the phenomena of fake text reviews on digital platforms and how it has been addressed in tourism. They employed the PRISMA methodology and concluded that further research on alternative approaches’ impact is necessary to detect fake online reviews for the tourism business. [39] investigated cognitive bias in the tourist’s decision-making. Thirty-seven articles were reviewed.

This paper [17] presents an analysis of various machine learning algorithms such as Multinomial Naive Bayes, Random Forest Classifier, and Bernoulli's Naive Bayes, and examines their behavior. Additionally, the study also includes examination of deep learning algorithms like Convolutional Neural Networks and Recurrent Neural Networks to determine their performance. Using the results obtained, a recommendation system is built that maps an individual user's interests to the highest rated tourist places and generates a unique tour plan tailored to the user's needs. The study finds that the Recurrent Neural Network model yielded the highest accuracy of 94.56%. As a result, it was concluded that deep learning algorithm outperforms traditional machine learning algorithms and is chosen to classify the user reviews.

An IoT-enabled deep learning-based recommendation system is proposed [18] to enhance the tourist experience in a smart city. Tourists will enter personal information and preferences into the smart city app/website, and the system will recommend activities and attractions that best fit their profile. In real-time, IoT devices will gather context-related information such as location and weather to suggest additional activities and/or attractions. The proposed multi-label deep learning classifier outperforms other models and successfully recommends tourist attractions for both pre-travel planning (loss: 0.5%, accuracy: 99.7%, precision: 99.9%, recall: 99.9%, F1-score: 99.8%) and in-city activities (loss: 3.7%, accuracy: 99.5%, precision: 99.8%, recall: 99.7%, F1-score: 99.8%).

This study [19] presents a new product recommendation system that utilizes an autoencoder based on a collaborative filtering method. The performance of this model is evaluated against the Singular Value Decomposition technique and the results are provided in the results section. The experiment demonstrates a high level of accuracy, as the recommendations align with user interests and do not suffer from the issue of data sparsity. The Root Mean Squared Error (RMSE) is found to be very low, specifically 0.996, with the system achieving an RMSE value of 0.029 in the first dataset and 0.010 in the second one, indicating a promising performance.

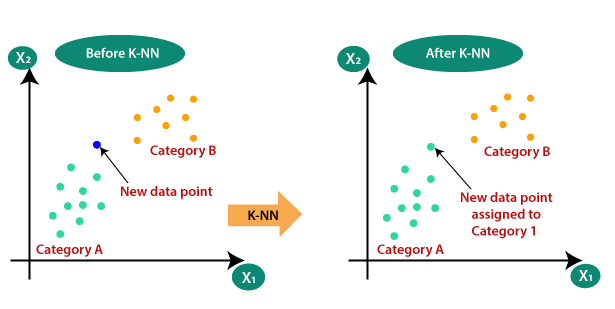


Figure 6: KNN algorithme.

# References

[1] Judith Hurwitz, Daniel Kirsch, and John Wiley. Machine Learning Machine Learning For Dummies ® , IBM Limited Edition. 2008, p. 69. isbn: 9781119454953. url: <http://www.wiley.com/go/permissions>.

[2] Pieter Th. van. Reenen, Margot van. Mulken, and J. W. Dyk. Studies in stemmatology. John Benjamins Pub. Co, 1996, p. 311. isbn: 9027221537. url: https: //books.google.de/books/about/Studies\_in\_Stemmatology.html?id=eRk\_ XdVl\_xMC&redir\_esc=y.

[3] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. http: //www.deeplearningbook.org. MIT Press, 2016.

[4] Porter, M. F. (1980). An algorithm for suffix stripping. Program, 14(3), 130-137.

[5] Nivre, J. (2006). Inductive dependency parsing. Dissertations in computational linguistics, 3, 1-124.

[6] Brill, E. (1992). A simple rule-based part of speech tagger. In Proceedings of the third conference on Applied natural language processing (pp. 152-155).

[7] Church, K. W. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In Proceedings of the second conference on applied natural language processing (pp. 136-143).

[8] Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of the 2003 conference of the North American chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1 (pp. 173-180).

[9] Earley, J. (1970). An efficient context-free parsing algorithm. Communications of the ACM, 13(2), 94-102.

[10] Charniak, E. (2000). A maximum-entropy-inspired parser. In Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference (pp. 132-139).

[11] Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. In Proceedings of the 41st annual meeting on association for computational linguistics-Volume 1 (pp. 423-430).

[12] Kiperwasser, E., & Goldberg, Y. (2016). Simple and accurate dependency parsing using bidirectional LSTM feature representations. Transactions of the Association for Computational Linguistics, 4, 313-327.

[13] Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticæ Investigationes, 30(1), 3-26.

[14] Ratinov, L., & Roth, D. (2009). Design challenges and misconceptions in named entity recognition. Proceedings of the Language Resources and Evaluation Conference, (May), 479-486.

[15]Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. Proceedings of the NAACL-HLT, 260-270.

[16] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[17] KULKARNI, Abhishek, SAMANT, R. M., BARVE, Prathamesh, *et al.* A Machine Learning Approach to Building a Tourism Recommendation System using Sentiment Analysis. *International Journal of Computer Applications*, vol. 975, p. 8887.

[18] Cepeda-Pacheco, J.C., Domingo, M.C. Deep learning and Internet of Things for tourist attraction recommendations in smart cities. *Neural Comput & Applic* **34**, 7691–7709 (2022). <https://doi.org/10.1007/s00521-021-06872-0>

[19] Ferreira, D.; Silva, S.; Abelha, A.; Machado, J. Recommendation System Using Autoencoders. Appl. Sci. **2020**, 10, 5510. <https://doi.org/10.3390/app10165510>

[20] ARTEMENKO, Olga, PASICHNYK, Volodymyr, KUNANETS, Nataliia, et al. Using sentiment text analysis of user reviews in social media for e-tourism mobile recommender systems. In : COLINS. 2020. p. 259-271.

[21] Uppink Calderwood, L., Soshkin, M., Jul. 2022. The travel and tourism competitiveness report 2019. [Online; accessed 1. Jul. 2022].

[22] F.A.C. Calderón, M.V.V. Blanco Impacto de internet en el sector turístico Revista UNIANDES Episteme, 4 (4) (2017), pp. 477-490

[23] Alvarez-Carmona, M. Á., Aranda, R., et al., 2022. Determinación automática del color del semáforo mexicano del covid-19 a partir de las noticias.

[24] Marr, B., 2016. How big data and analytics are changing hotels and the hospitality industry. Forbes. com, January 26.

[25] K. Gupta, T. Gauba, S. Jain

Big data in hospitality industry: A survey Int. Res. J. Eng. Technol., 4 (11) (2017), pp. 476-479

[26] S. Nadkarni, F. Kriechbaumer, M. Rothenberger, N. Christodoulidou The path to the hotel of things: Internet of things and big data converging in hospitality J. Hospitality Tourism Technol. (2019)

[27] P.M. Nadkarni, L. Ohno-Machado, W.W. Chapman Natural language processing: an introduction J. Am. Med. Inform. Assoc., 18 (5) (2011), pp. 544-551

[28] T. Cai, A.A. Giannopoulos, S. Yu, T. Kelil, B. Ripley, K.K. Kumamaru, F.J. Rybicki, D. Mitsouras Natural language processing technologies in radiology research and clinical applications Radiographics, 36 (1) (2016), pp. 176-191.

[29] Putri, F.P., Meidia, H., Gunawan, D., 2019. Designing intelligent personalized chatbot for hotel services. In: Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence. pp. 468–472.

[30] Bulchand-Gidumal, J., 2020. Impact of artificial intelligence in travel, tourism, and hospitality. Handbook of e-Tourism; Xiang, Z., Fuchs, M., Gretzel, U., Höpken, W., Eds, 1–20.

[31] O. Loyola-González, M.A. Medina-Pérez, K.-K.R. Choo

A review of supervised classification based on contrast patterns: Applications, trends, and challenges

J. Grid Comput. (2020), pp. 1-49

[32] J. Rodríguez-Ruiz, M.A. Medina-Pérez, R. Monroy, O. Loyola-González

A survey on minutiae-based palmprint feature representations, and a full analysis of palmprint feature representation role in latent identification performance

Expert Syst. Appl., 131 (2019), pp. 30-44

[33] Close M. Weed Sports tourism research 2000–2004: A systematic review of knowledge and a meta-evaluation of methods

Journal of Sport & Tourism, 11 (1) (2006), pp. 5-30

[34] L.G. Anderson, S. Rocliffe, N.R. Haddaway, A.M. Dunn

The role of tourism and recreation in the spread of non-native species: a systematic review and meta-analysis

PloS one, 10 (10) (2015), Article e0140833

[35] D.O. Gomezelj

A systematic review of research on innovation in hospitality and tourism Int. J. Contemporary Hospitality Manage. (2016)

[36] J. Li, L. Xu, L. Tang, S. Wang, L. Li Big data in tourism research: A literature review

Tourism Manage., 68 (2018), pp. 301-323

[37] A. Marasco, M. De Martino, F. Magnotti, A. Morvillo Collaborative innovation in tourism and hospitality: a systematic review of the literature Int. J. Contemporary Hospitality Manage. (2018)

[38] A. Reyes-Menendez, J.R. Saura, F. Filipe The importance of behavioral data to identify online fake reviews for tourism businesses: A systematic review PeerJ Computer Science, 5 (2019), Article e219

[39] W. Wattanacharoensil, D. La-ornual A systematic review of cognitive biases in tourist decisions

Tourism Management, 75 (2019), pp. 353-369

[40] BREIMAN, Leo. Random forests. *Machine learning*, 2001, vol. 45, no 1, p. 5-32.

[41] ROKACH, Lior et MAIMON, Oded. Decision trees. In : *Data mining and knowledge discovery handbook*. Springer, Boston, MA, 2005. p. 165-192.

[42] QUINLAN, J.. Ross . Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, 1996, vol. 28, no 1, p. 71-72.

[43] JENA, Monalisa et DEHURI, Satchidananda. DecisionTree for Classiﬁcation and Regression: A State-of-the Art Review. *Informatica*, 2020, vol. 44, no 4.

[44] Breiman, L., & Cutler, A. (2016). Random Forests for scientific discovery. *línea]. Available: https://www. stat. berkeley. edu/~ breiman/RandomForests/berkeley\_files/frame. htm*.‏

[45] CUNNINGHAM, Pádraig, CORD, Matthieu, et DELANY, Sarah Jane. Supervised learning. In : *Machine learning techniques for multimedia*. Springer, Berlin, Heidelberg, 2008. p. 21-49.

[46]R. Schwitter, “COMP8220: Machine Learning What is Machine Learning?,” 2020.

[47] G. Shaheamlung, H. Kaur, and M. Kaur, “A Survey on machine learning techniques for the diagnosis of liver disease,” Proc. Int. Conf. Intell. Eng. Manag. ICIEM 2020, pp. 337–341, 2020.

[48] ZHU, Xiaojin et GOLDBERG, Andrew B. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 2009, vol. 3, no 1, p. 1-130.

[49] ZHU, Xiaojin Jerry. Semi-supervised learning literature survey. 2005.