An intelligent method of detecting and classifying hotels for an efficient choice

Mamadou Diarra, Issiaka Sanou, Abdoulaye Sere

An intelligent method of detecting and classifying hotels for an efficient choice

Mamadou DIARRA[*1], **Issiaka SANOU**[2], **Abdoulaye SERE**[2]

[1]Ecole Polytechnique de Ouagadougou, Burkina Faso
[2]Université Nazi Boni, Burkina Faso

*E-mail : diarra.md21@gmail.com

Abstract

The hotel industry attaches a great deal of importance to customer satisfaction and loyalty by offering quality services. However, given the security situation in Burkina Faso, the proximity of administrative and public services, health services and the quality of access roads have become other essential criteria in the choice of accommodation sites. Given that manual analysis of such a large number of criteria and the volume of data generated proves utopian, the search for solutions to aid decision-making is imperative.

To address this concern, we opted for a comparative study of machine learning and multi-criteria optimisation for better prediction of customer choice. We also carried out a comparative study of different classification techniques to determine the best approach.

We then used the correlation matrix to determine the coefficients applied in the Pareto principle. The different approaches enabled us to confirm that the Extra Trees Regressor gave a better prediction of the choice of hoteliers. We then determined the coefficients of the endogenous criteria in our study using a correlation matrix. From the application of the Pareto principle, we can affirm that: security, pharmacies and health centres are the determining factors in the choice of hotel complexes.

Keywords

Hotel Market, External Environmental Factors, Machine Learning, Multi-Criteria Optimisation

I INTRODUCTION

Tourism continues to flourish in Burkina Faso, despite the difficult security situation. The organisation of major international events such as National Culture Week (SNC), Ouagadougou International Crafts Fair (SIAO), Tour of Faso, Ouagadougou Pan-African Film and Television Festival (FESPACO), Ouagadougou international tourism and hotel trade fair (SITHO) and many others are a perfect illustration of this. These events are a focal point for the populations of Burkina Faso and other countries, creating a need for comfort and security.

In an era of digital revolution, increasingly data-driven, the benefits of machine learning are clear to see. In an ever-changing business environment, where every detail counts in the decision-making process, machine learning algorithms can provide a strategic advantage that can make all the difference.

In this way, the prediction of certain endogenous factors such as comfort, room type and overnight stay, and exogenous factors such as airport and train stations, security services and hospitals and pharmacies, relating to hotel sites will help to influence customer choice. To predict this choice, we are carrying out a comparative study between Machine Learning [1], which refers to a set of inductive processes whose objective is to learn from data, and Multi-Criteria Optimisation [2], an approach that enables the best possible solution to be found by optimising a set of criteria and constraints.

This work is organised into three main parts. First, we present the state of the art, then we develop our contribution through the prediction methods chosen, and finally the evaluation and results.

II RELATED WORKS

2.1 Hotel market and external environmental factors

The hotel industry includes all activities related to traditional hotels and large hotel groups, with the provision of short-stay accommodation for business travellers (and even other groups), as well as accommodation facilities for longer or shorter stays [3]. External environmental factors represent characteristics or trends in the company's environment over which it has no control.

In this study, the factors concerned were safety, health, accessibility and distance from the hosting site.

2.2 Machine Learning, multi-criteria optimisation

2.2.1 Machine Learning Concepts

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy [4] (Figure 1). Moreover, machine learning focuses on discovering correlation between data elements, recognizing data patterns, and performing tasks without additional human instructions. Because machine learning often uses an iterative approach to learn from data, the learning routines and processes can be easily automated.

Fundamentally, machine learning is focused on the analysis of data for structure, even if the structure is not known ahead of time. Moreover, machine learning is focused on the implementation of computer programs and systems which can teach themselves to adapt and evolve when introduced to new data. At the core of machine learning are computer algorithms, which are procedures for solving a mathematical problem in a finite number of steps. And machine learning algorithms are utilized to build a mathematical model of sample data, known as "training data".

2.2.2 Machine Learning Process

To carry out this comparative study of the performance between machine learning and multi-criteria optimisation in predicting the choice of a hotel establishment, we proceed as follows:

- Acquire dataset / data source: Dataset is a collection of related, discrete items of related data that may be accessed individually or in combination or managed as a whole entity. It is organized into some type of data structure. A dataset is a collection of data objects referred to as points, patterns, events, cases, samples, observations, or entities [9]

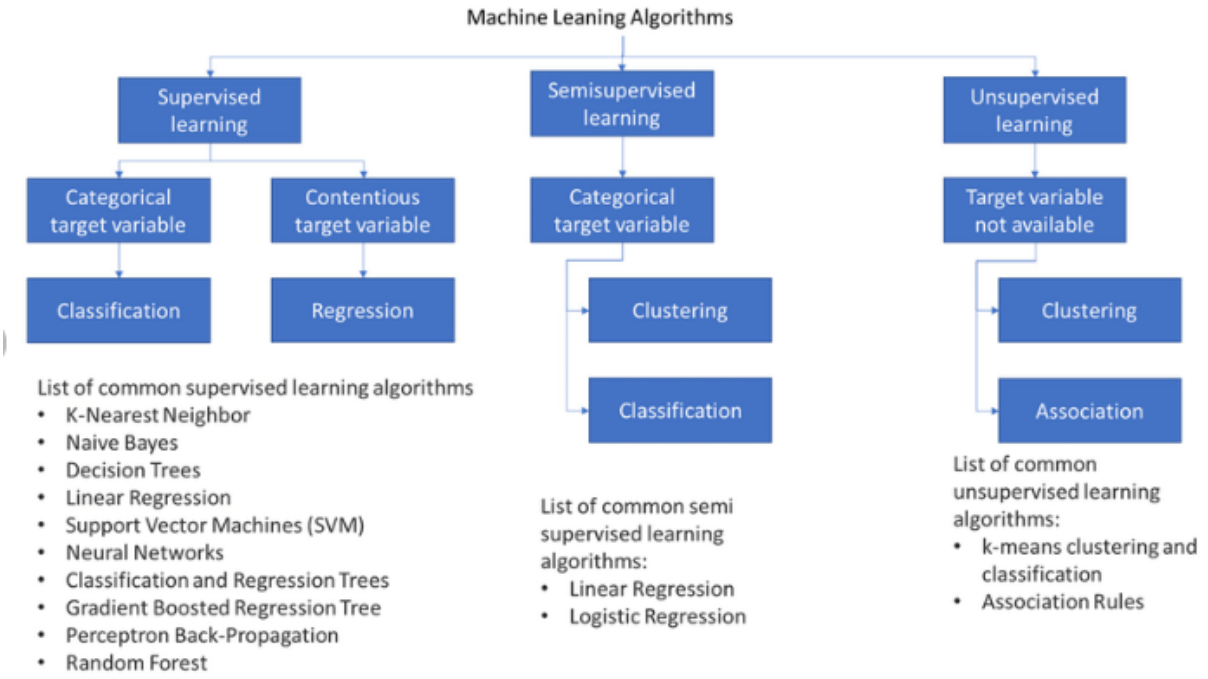

Figure 1: Machine learning algorithms classification. [6]

- Data preparation / cleansing and applying Exploratory Data Analysis (EDA) on data such as investigating Null data, dummy variables, under sampling / over sampling and Correlation among all the variables in the dataset.
- Create training and test datasets and evaluate the datasets to make sure we have done sampling correctly (Figure 2). The training data is the biggest (in -size) subset of the original dataset, which is used to train or fit the machine learning model. The test dataset is another subset of original data, which is independent of the training dataset.
- Apply / Deploy ML model. Once an ML model is ready, the next step is to make it available to users so that they can make predictions. Making an ML model available in a business environment is called deploying the model (Figure 3).
  For example, from the airport, a customer can use a model to predict a score of hotels where he or she can stay, depending on an event, using a set of criteria. Predicting the choice can be as simple as calling this function: Prediction = classifier.predict(INPUT DATASET).
  In predicting hotel choice, we were interested in 13 categorical variables and classification, which is one of the supervised learning methods used to return a discrete value to be explained, such as a class or label. In this way, the model is trained to output a discrete value (y) as a function of an input value (x). The main classification algorithms are listed in Figure 1.
- Evaluate deployed models To monitor and evaluate your models, you need to collect and store data from various sources, such as inputs, outputs, predictions, errors, logs, or feedback. You should design a data pipeline that can handle the volume, variety, and velocity of your data, and ensure its quality, security, and accessibility.

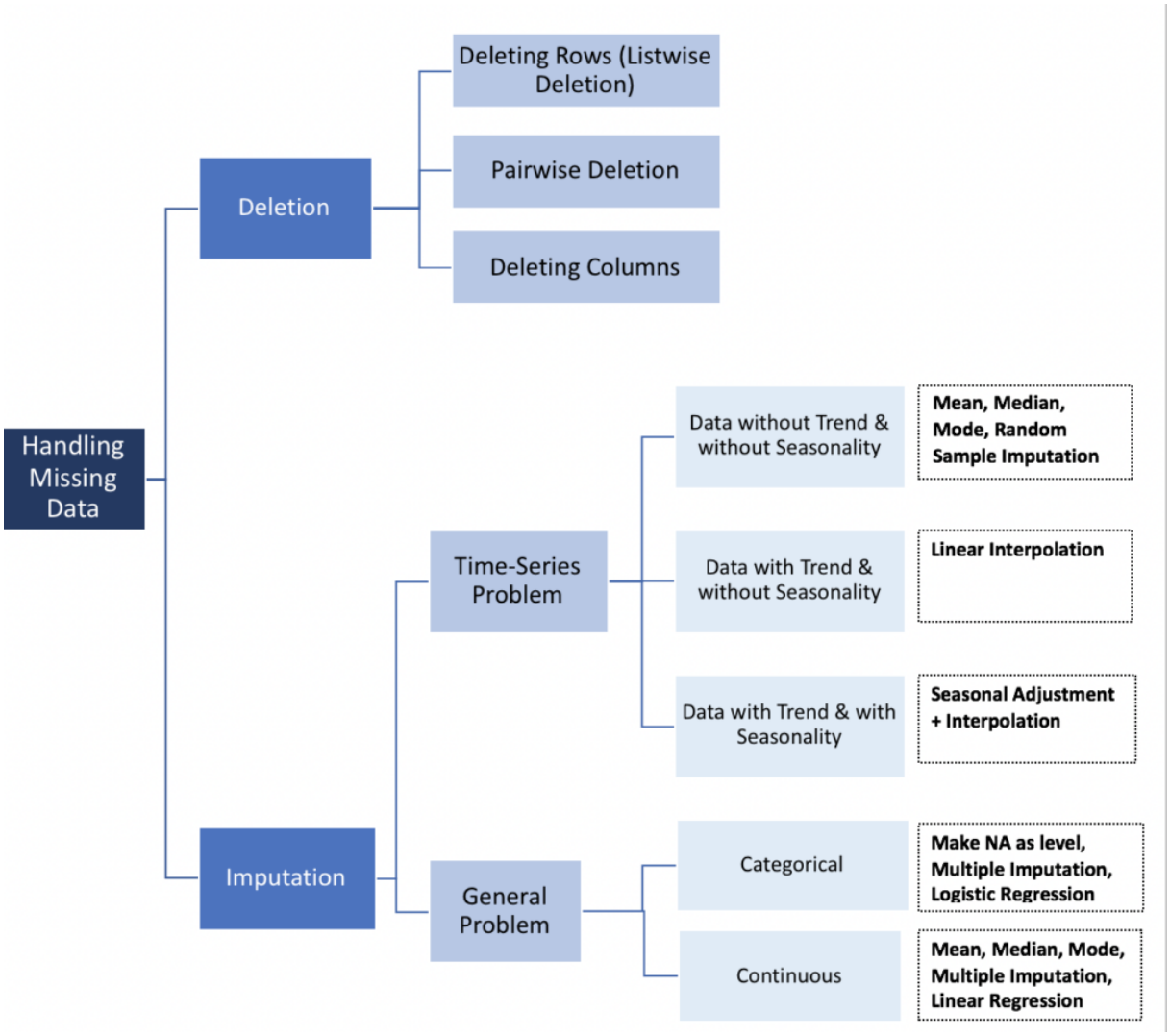

Figure 2: How to manage missing data. [11]

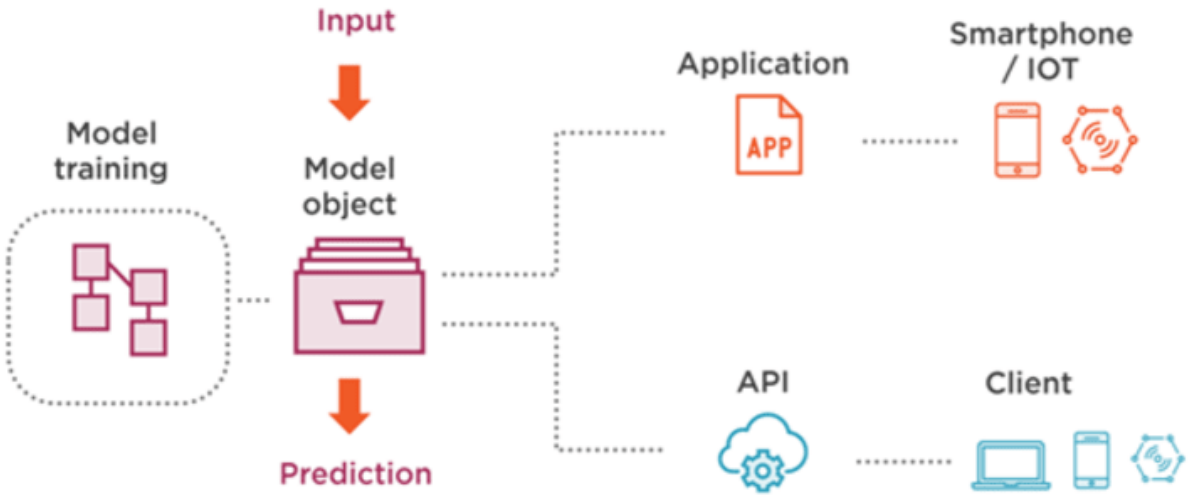

Figure 3: From training to model serving [10]

2.3 Multi-criteria optimisation

In general, multi-criteria optimization [5, 7, 8] (Figure 4), requires more computational effort than single-criteria optimization. Methods with a priori articulation of preferences require the

user to specify preferences only in terms of objective functions. Alternatively, methods with a posteriori articulation of preferences allow the user to view potential solutions in the criterion space [7].

Selection of a specific scalarization method for a priori articulation of preferences which allows the user to design a utility function depending on the type of preferences that the decision maker wishes to articulate.

Most multi-criteria optimization algorithms depend on the efficiency of the single-objective optimization algorithm. Hence, it is necessary to select an efficient single-objective optimization algorithm and associated software [7, 8].

Moreover, methods with a posteriori articulation of preference are less efficient than methods with a priori articulation of preferences in terms of computer process unit time.

Though, this paper has not exhausted the multi-criteria optimization methods. Many other multi-criteria methods such as reference point method, genetic algorithm, exponential weighted, bounded objective function, etc. may also be useful [5].

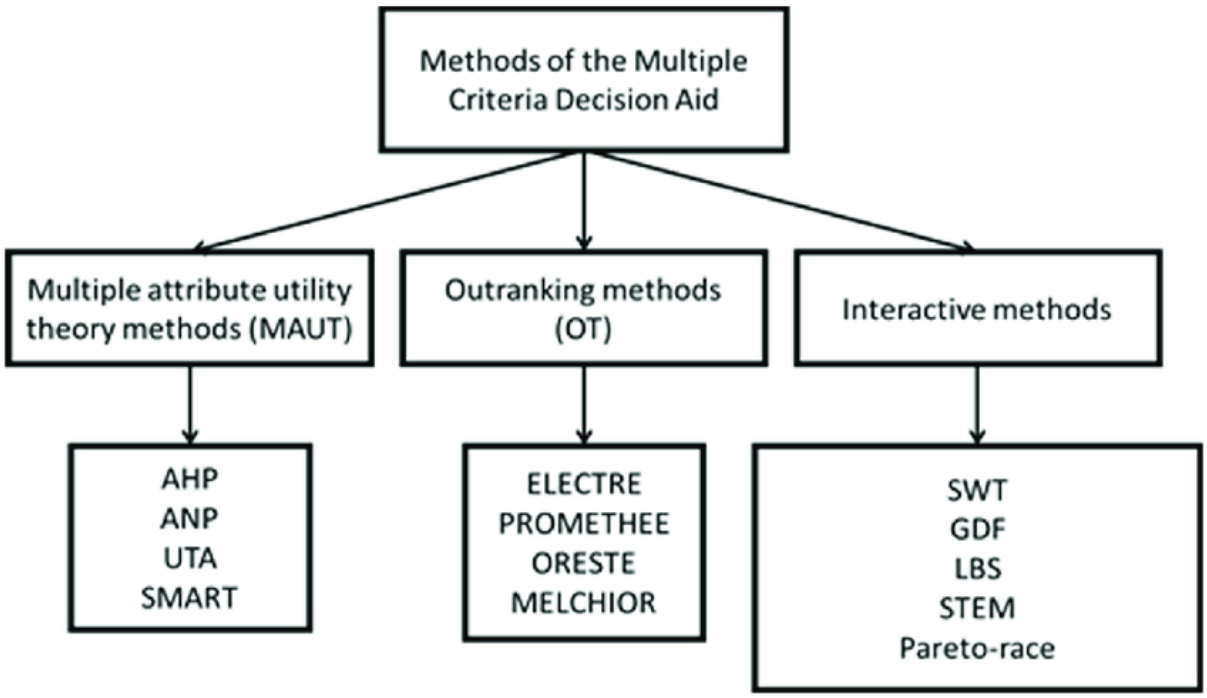

Figure 4: Classification of multi-criteria decision aid (MCDA) methods according to P. Vincke. [8]

III OUR CONTRIBUTIONS

3.1 Methodologies

To carry out this comparative study of the performance of machine learning and multi-criteria optimisation in predicting the choice of a hotel establishment, we proceeded as follows:

3.1.1 Acquire dataset or data source

Our data comes from the study: "Market orientation and commercial performance of hotel establishments in the cities of Ouagadougou and Bobo-Dioulasso: the moderating role of the external environment". The initial collection consisted of 47 criteria and 500 records.

For our study, we focused on 7 external criteria . also, we found that the survey took into account the 13 regional capitals of Burkina Faso, which justifies the number of registrations .

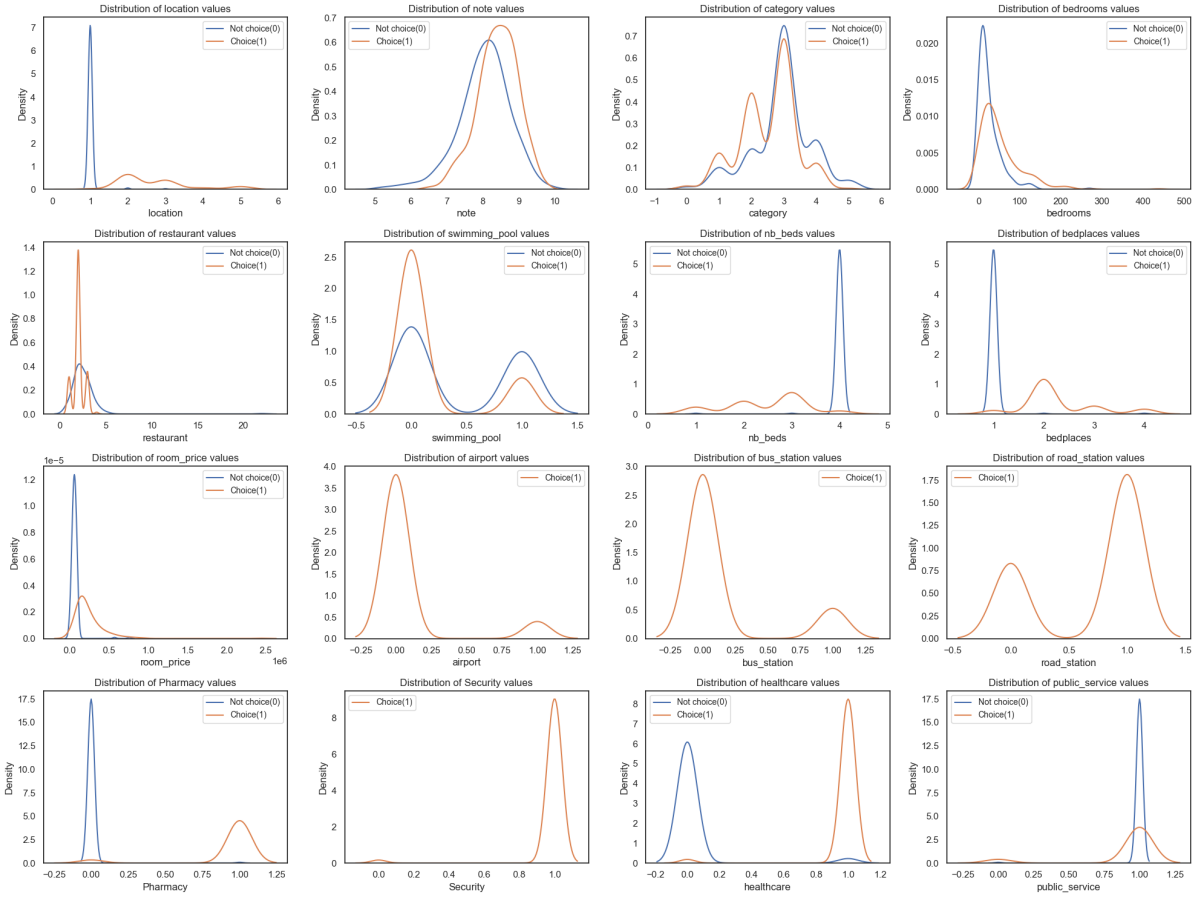

Figure 5: Distribution of features

3.1.2 Data preparation/cleansing and applying Exploratory Data Analysis (EDA)

It consisted in extracting 13 criteria which were the subject of this study, namely 6 categorical criteria related to comfort and the 7 external environmental criteria.

Integrating huge amounts of variable data requires various strategies and resources to homogenise them. To ensure that all the data is of the same quality, it must also be cleaned and filtered before being converted and integrated. To do this, we explored the data before correcting missing, erroneous and/or outlier data using the median method. We then labelled the variables. Finally, we standardised the data to improve its efficiency and ensure that it corresponded to a predefined and constrained set of values without distorting it. For numerical data, we used the min-max method or z-score normalisation, and one-hot encoding to transform categorical data into binary variables.

For the 7 external factors, we assigned a score of 1 to external factors located more than 2 km from the hotels and 0 in the opposite case (Figure 6).

	Security	airport	bus_station	road_station	Pharmacy	healthcare	public_service
count	500.000000	500.000000	500.000000	500.000000	500.000000	500.000000	500.000000
mean	0.544000	0.052000	0.086000	0.380000	0.518000	0.558000	0.946000
std	0.498559	0.222249	0.280645	0.485873	0.500176	0.497122	0.226244
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
50%	1.000000	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000
75%	1.000000	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Figure 6: Target data description

3.1.3 Create training and test datasets

For the simulation, we used a PC running Windows 10, with an Intel(R) Core(TM) i5-6200U @ 2.30 GHz 2.40 GHz processor and 16 GB RAM. The 500 recordings were divided into training and test data with a proportion of 80/20 or 400 recordings for training and 100 for test. This choice was made on the basis of the Pareto principle. The target values are the following external factors: security, airport, bus station, pharmacy, healthcare and public services. Also, a correlation matrix will allow us to determine the weights of the factors for the multi-criteria optimisation (Figure 7).

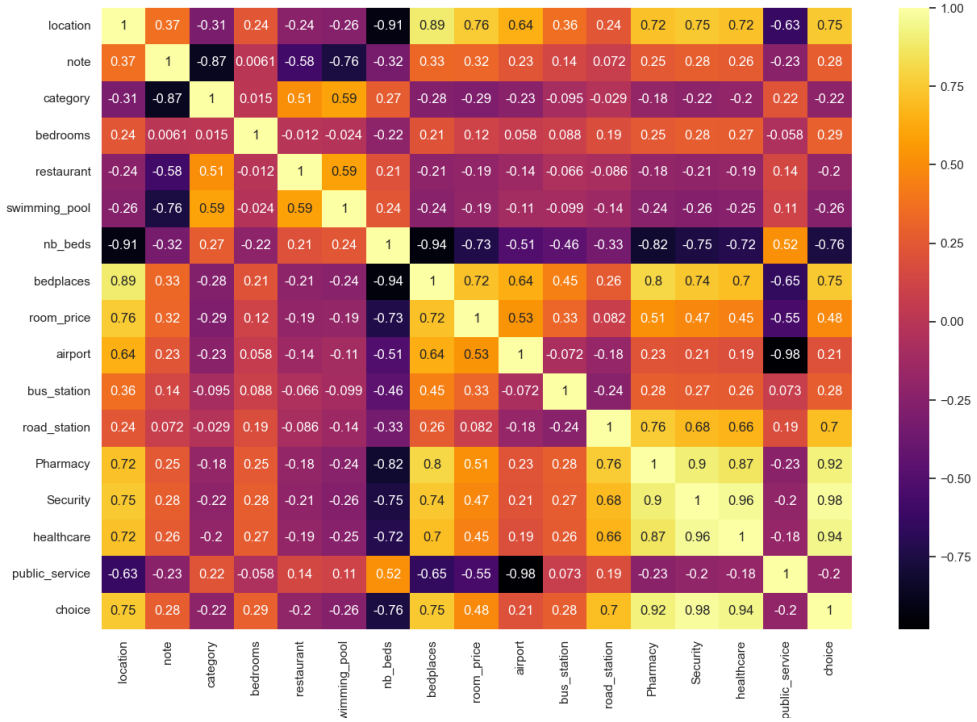

Figure 7: Correlation matrix for dataset

3.2 Applying machine learning

3.2.1 Study results

Once the data preparation process was complete, we extracted a different dataset for each target variable, before splitting them into a training set and a validation set in an 80/20 proportion (figure 8). We then trained 5 models on each training set before measuring the RMSE [12] and R2-Score [13] . Given the wide range of values taken by the choice variable, in this case the model was trained to predict the logarithmic value of the target column. The scores obtained are shown in the table 1.

Model		MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
et	Extra Trees Regressor	0.0467	**0.0213**	**0.1431**	**0.9122**	**0.0976**	0.0451	0.5380
rf	Random Forest Regressor	0.0516	0.0285	0.1648	0.8829	0.1138	0.0475	0.7240
catboost	CatBoost Regressor	0.0624	0.0281	0.1652	0.8847	0.1124	0.0651	2.7780
xgboost	Extreme Gradient Boosting	**0.044**9	0.0372	0.1883	0.8464	0.1298	**0.0334**	0.5900
lightgbm	Light Gradient Boosting Machine	0.0939	0.0376	0.1910	0.8459	0.1323	0.0947	0.6000
ada	AdaBoost Regressor	0.0929	0.0411	0.1991	0.8320	0.1449	0.0559	0.4940
gbr	Gradient Boosting Regressor	0.0654	0.0451	0.2098	0.8146	0.1408	0.0601	0.4860
dt	Decision Tree Regressor	0.0480	0.0480	0.2113	0.8026	0.1465	0.0483	0.5760
knn	K Neighbors Regressor	0.1168	0.0731	0.2625	0.7015	0.1842	0.1068	0.4840
en	Elastic Net	0.2853	0.1367	0.3643	0.4419	0.2292	0.2900	0.4540
ridge	Ridge Regression	0.2570	0.1440	0.3699	0.4121	0.2157	0.2720	0.4820
br	Bayesian Ridge	0.2579	0.1444	0.3706	0.4107	0.2167	0.2718	0.5460
lasso	Lasso Regression	0.3224	0.1485	0.3811	0.3934	0.2535	0.3193	0.5280
llar	Lasso Least Angle Regression	0.3224	0.1485	0.3811	0.3934	0.2535	0.3193	0.5340
lr	Linear Regression	0.2606	0.1581	0.3814	0.3550	0.2216	0.2722	4.7900
lar	Least Angle Regression	0.2606	0.1581	0.3814	0.3550	0.2216	0.2722	0.4700
huber	Huber Regressor	0.2395	0.1746	0.4016	0.2875	0.2108	0.2841	0.4740
omp	Orthogonal Matching Pursuit	0.3743	0.1889	0.4319	0.2276	0.2882	0.3568	0.4900
dummy	Dummy Regressor	0.4951	0.2472	0.4972	-0.0106	0.3473	0.4547	0.3880
par	Passive Aggressive Regressor	0.4119	0.5141	0.6881	-1.1157	0.2527	0.5460	0.4720

Table 1: Classification models comparative study

3.2.2 Analysis of results

For the evaluation of the 20 classification models, we carried out a comparative study of 6 evaluation metrics MAE, MSE, RMSE, R2, RMSLE, MAPE and the execution time TT (Sec) of the models. The comparative study of the metrics MAE, MSE, RMSE, R2, RMSLE, MAPE, and the execution time TT (Sec) 0.0467 0.0213 0.1431 0.9122 0.0976 0.0451 0.5380

It should be noted that despite the undeniable performance of the Extra Trees Regressor on all metrics. Extreme Gradient Boosting however has a better performance compared to the Extra Trees Regressor regarding the MAE metrics with score of 0.0449 and MAPE with score of 0.0334 which are significantly lower than 0.0467 and 0.0451

3.3 Applying multi-criteria optimisation

3.3.1 Study results

Application of the Pareto principle, also known as the 80/20 rule, based on 13 categorical criteria and 500 observations. We used the correlation matrix (CM : Figure 7) and principal component analysis (PCA). These different approaches enabled us to weight the different criteria. The results of the 2 approaches give us the table 2:

We then applied Pareto's 80/20 principle to the different weights obtained with the correlation matrix, as both methods provide almost the same weighting. This approach produced the following table 3 and graph 8

Features	CM weight	ACP weight
location	75	75
note	28	27.7
category	22	21.8
restaurant	29	21.9
airport	21	21
bus_station	28	27.5
road_station	70	70.2
Pharmacy	92	92.2
Security	98	98
healthcare	94	94.3
public_service	20	23

Table 2: Multicriteria optimisation: optimality of Pareto sense

N°	*Features*	*Count*	*count cumsum*	*cumpercentage*
8	Security	98	98	0.169844
9	healthcare	94	192	0.332756
7	Pharmacy	92	284	0.492201
0	location	75	359	0.622184
6	road_station	70	429	0.743501
3	restaurant	29	458	0.793761
1	note	28	486	0.842288
5	bus_station	28	514	0.890815
2	category	22	536	0.928943
4	airport	21	557	0.965338
10	public_service	20	577	1.000000

Table 3: Multicriteria optimisation: optimality of Pareto sense

3.3.2 Analysis of results

The application of the Pareto principle, on the various environmental factors attests that only 3 of the 7 have an influence on the choice of hotels. This is the highest level of safety at 98 followed by health centres at 92 and pharmacies at 92. Also, it interpretation of the law of Pareto, allows us to affirm according to the curve decision (Figure 8) that its 3 criteria represent 80% of the decisions of choice of hotels.

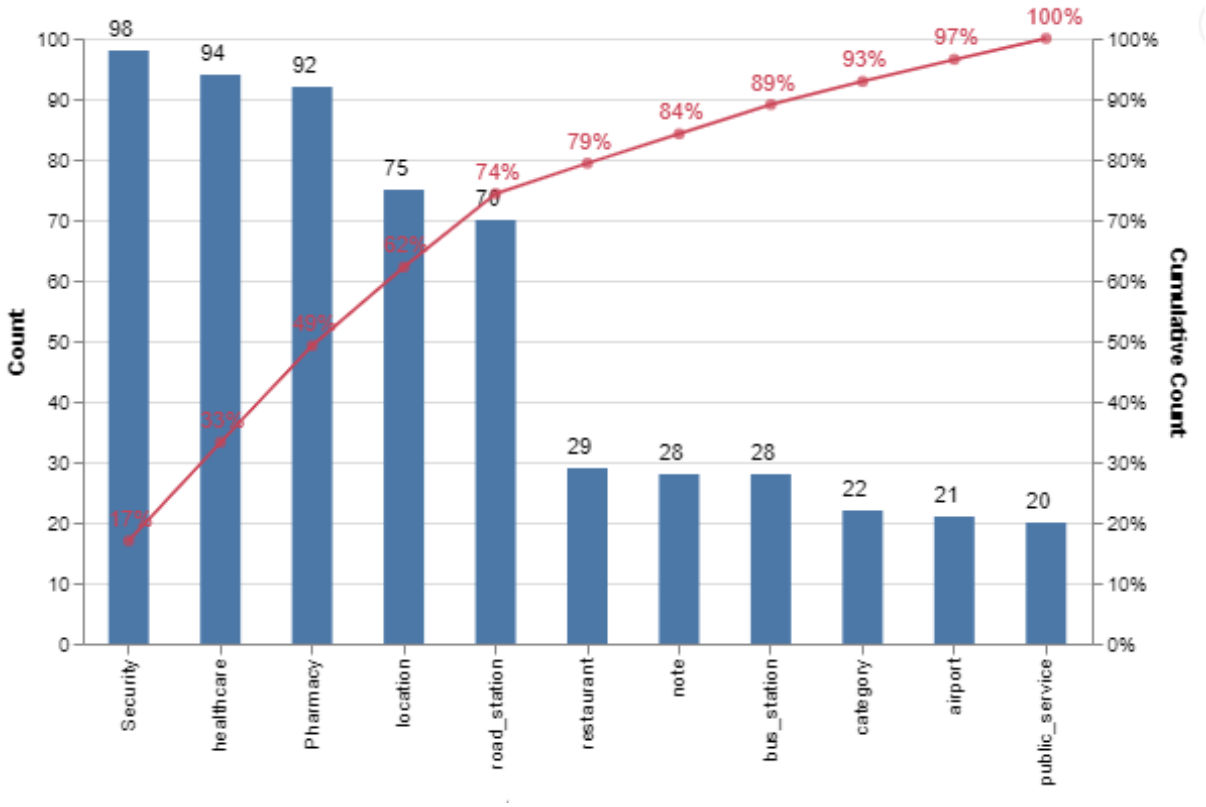

Figure 8: Pareto chart of customer priorities

IV CONCLUSION AND OUTLOOKS

Our study highlights the link between market orientation and the impact of environmental factors, with more than 90% prediction of customer choice. The comparative study of different classification models attests that the Extra Trees Regressor is the method that gives the best prediction of the customer's choice. Indeed, it gets the lowest RMSE (0.1431) and the highest R2-Score (0.9122).

Also, the application of the Pareto principle highlights 3 factors determining the choice of customers namely safety, centers and pharmacies which 20% cause causing 80% consequences.

Thus, obtaining its different information through this study confirms that machine learning can predict customer choice, but multi-criteria optimization can reduce the number of determining factors for prediction.

In perspective, we will make a comparative study of the prediction of the criteria before and after optimization to check the quality of the prediction.

REFERENCES

[1] BOILEAU, J. É., Bois-Drivet, I., Westermann, H., & Zhu, L. (2022). Rapport sur l'épistémologie de l'intelligence artificielle (IA). Laboratoire de cyberjustice.

[2] Belmecheri, N., Aribi, N., Lazaar, N., Lebbah, Y., & Loudni, S. (2022, January). Une méthode d'apprentissage par optimisation multicritère pour le rangement de motifs en fouille de données. In ECG 2022-Extraction et Gestion des Connaissances.

[3] Umaya, S., & Syawalina, L. Analyse des Besoins des Ètudiants pour le Cours de Français de l'Hôtellerie et de la Restauration. HEXAGONE Jurnal Pendidikan, Linguistik, Budaya dan Sastra Perancis, 12(1), 16-25.

[4] https://www.ibm.com/topics/machine-learning 20.08.2023

[5] Odu, G. O., & Charles-Owaba, O. E. (2013). Review of multi-criteria optimization methods–theory and applications. IOSR Journal of Engineering, 3(10), 01-14.

[6] Aldahiri, Amani & Alrashed, Bashair & Hussain, Walayat. (2021). Trends in Using IoT with Machine Learning in Health Prediction System. Forecasting. 3. 181-207. 10.3390/forecast3010012.

[7] Marttunen, M. (2010). Description of multi-criteria decision analysis (mcda). Finnish Environment Institute.

[8] Nosal, Katarzyna & Solecka, Katarzyna & Szarata, Andrzej. (2019). The Application of the Multiple Criteria Decision Aid to Assess Transport Policy Measures Focusing on Innovation. Sustainability. 11. 1472. 10.3390/su11051472.

[9] Karuppusamy, P., Perikos, I., Shi, F., & Nguyen, T. N. (2020). Sustainable communication networks and application. Lecture Notes on Data Engineering and Communications Technologies, 65-72.

[10] Vasconcelos, R. (2021, 17 mai). A guide to ML model serving. Ubuntu. Consulté le 4 juin 2023, à l'adresse https://ubuntu.com/blog/guide-to-ml-model-serving

[11] Guerard_guillaume. (2022, 24 avril). Comment gérer les données manquantes - Complex systems and AI. Complex systems and AI. https://complex-systems-ai.com/analyse-des-donnees/comment-gerer-les-donnees-manquantes/

[12] Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. Geoscientific model development, 7(3), 1247-1250.

[13] Lakshminarayanan, S. K., & McCrae, J. P. (2019, December). A Comparative Study of SVM and LSTM Deep Learning Algorithms for Stock Market Prediction. In AICS (pp. 446-457).