Artificial Intelligence Applied to Address Tourism Challenges: Predicting Hotel Room Cancellations

International Conference on Emerging Challenges:

Smart Business and Digital Economy

Ngô-Hồ Anh-Khôi[1*], Lục Hà-Duy-Nguyên[2], and Triệu Vĩnh-Khang[1]

[1] Nam Can Tho University, Can Tho City, Vietnam

[2] Van Lang University, Ho Chi Minh City, Vietnam

[*]Corresponding author: ngohoanhkhoi@gmail.com

Abstract

The contemporary era has witnessed a substantial surge in the application of artificial intelligence across diverse domains of research. Notably, the integration of machine learning techniques has garnered considerable attention in the realm of forecasting economic phenomena. Of particular interest in recent times is the predictive modeling of hotel room cancellations, an issue for which conventional research methodologies have proven inadequate. The intricacies of this problem demand sophisticated predictive capabilities that are best addressed through the deployment of artificial intelligence.This study is centered on the utilization of continuous learning algorithms with the primary objective of harnessing existing datasets while accommodating evolving prediction requirements. The methodology is designed to adapt progressively to the distinct characteristics of Vietnamese data, thereby ensuring robust predictive performance. The core innovation in this research lies in the amalgamation of the sliding window approach within the framework of continuous learning, coupled with the selection of classical machine learning algorithms in artificial intelligence. This amalgamation transforms the selected classical algorithms into continuous learning models, custom-tailored to the specific demands of this economic phenomena. The study's findings are systematically juxtaposed to discern the optimal algorithm for forecasting hotel room cancellations—an issue of substantial economic significance, especially within the context of Vietnam. To facilitate both economic researchers' experimentation and the practical implementation of this methodology within hotel establishments, a dedicated website has been established. This platform serves as a valuable resource for evaluating the real-world utility of the proposed approach.

Research purpose:
The significance of this matter within the Vietnamese context is profound. Tourism plays a pivotal role in Vietnam's economic landscape, contributing significantly to the nation's Gross Domestic Product (GDP). Addressing the challenge of booking cancellations is of paramount importance, as it offers numerous advantages for both travelers and hoteliers. Effectively addressing this issue fosters trust among visitors and ensures the sustainable growth of the tourism sector in the long term, particularly in Vietnam.

Research motivation:
The issue of hotel booking cancellations presents a formidable challenge with far-reaching implications for both the global hospitality industry and travelers. Beyond its financial ramifications for hotels, cancellations significantly impact the quality of travelers' experiences. This challenge has been exacerbated during the Covid-19 pandemic and also after that. due to the heightened uncertainty stemming from the unpredictable nature of the outbreak and the subsequent implementation of social distancing policies. In the specific context of Vietnam, booking cancellations have emerged as a prominent issue.

Research design, approach, and method:
This study is centered on an emerging research area that aligns with an innovative algorithmic framework. Its primary aim is to contribute to solving the challenge of hotel room cancellations. To achieve this objective, the study relies on a meticulously curated database and is focused on identifying a suitable method within this dataset to serve as the foundational algorithm for a classification system. The ultimate goal is to provide organizations with the necessary tools, research insights, and practical outcomes to effectively address the issue of room cancellations through the utilization of an artificial intelligence system. This, in turn, will enhance the reliability of room reservations for both domestic and international establishments, not only in Vietnam but also in other regions. The research endeavor seeks to leverage

N. D. Nguyen and P. T. T. Hong (eds.), *Proceedings of the 11th International Conference on Emerging Challenges: Smart Business and Digital Economy 2023 (ICECH 2023)*, Advances in Economics, Business and Management Research 274,
https://doi.org/10.2991/978-94-6463-348-1_33

artificial intelligence systems to predict customer room booking needs. This approach aims to improve businesses' understanding of tourist requirements, ultimately leading to increased revenue and fostering economic progress, both within Vietnam and on a global scale.

Main findings:
The primary objective of this investigation is to conduct a comparative analysis of significant machine learning classifiers within the framework of evolving learning methodologies, contrasting them with conventional static machine learning approaches observed in prior research. This decision arises from the imperative need to embrace a continuous machine learning framework as opposed to a static model. This necessity arises due to the sustained utilization of the hotel cancellations dataset for future research endeavors pertaining to hotel cancellations in Vietnam. We embark on an exploration of a diverse set of six distinct machine learning algorithms in artificial intelligence (MLPClassifier, KneighborsClassifier, Linear Discriminant Analysis, BernoulliNB Classifier, Decision Tree Classifier and GaussianNB Classifier), enhanced by the incorporation of progressive strategies inspired by the Klinkenberg concept, to dynamically determine the optimal window size. In summary, the DecisionTreeClassifier algorithm appears to be highly suitable for the hotel room reservation prediction problem and practical applications.

Practical/managerial implications:
The research introduces a demonstration system designed to familiarize researchers with the practical application of the developed algorithms and systems. This topic is poised for further development in the future, which may involve updating the dataset through survey methods and incorporating specialized knowledge to acquire the most standardized and closely aligned dataset with reality. The system's functionalities have been thoroughly implemented, effectively meeting the initial requirements. It is designed with separate sections for general users and developers, offering a user-friendly interface for predicting hotel reservations while providing developers with additional pages for system enhancement and development.

Keywords: *hotel reservations dataset, hotel room cancellation prediction, sliding windows technique, machine learning in economics, AI application*

1. INTRODUCTION

Approximately half a decade ago, the integration of machine learning into the field of economics began to experience a flourishing expansion. It is noteworthy that the concept of digital economics was originally introduced in the 1990s by Tapscott and Don (Tapscott and Don, 1996); however, the predictive capabilities of this discipline at that time bore little resemblance to the sophisticated achievements realized today through the fusion of economics and machine learning. The historical origin of this amalgamation is challenging to trace accurately. Scant literature exists that explicitly delineates the outcomes of applying machine learning to economic forecasting before the year 2015, as evidenced by studies conducted by Kleinberg et al. (Kleinberg et al., 2015) or Bjorkegren and Grissen (Bjorkegren and Grissen, 2015). Nevertheless, it is important to emphasize that the developments witnessed thus far represent only a fraction of the potential contributions that machine learning can make to the field of economic research, as succinctly articulated by Susan Athey (Susan Athey, 2018). In concurrence with this prevailing trajectory, the research group endeavors to explore further a recent avenue of investigation concerning the prediction of hotel reservation cancellations. To achieve this, a combination of machine learning techniques is employed, aiming to unlock deeper insights into this specific economic phenomenon.

The issue of hotel booking cancellations presents a formidable challenge with far-reaching implications for both the global hospitality industry and travelers. Beyond its financial ramifications for hotels, cancellations significantly impact the quality of travelers' experiences. This challenge has been exacerbated during the Covid-19 pandemic due to the heightened uncertainty stemming from the unpredictable nature of the outbreak and the subsequent implementation of social distancing policies. In the specific context of Vietnam, booking cancellations have emerged as a prominent issue. As evidenced by a report from the Vietnam Tourism Association, the year 2020 witnessed a substantial increase in customer-initiated booking cancellations within the country's hospitality establishments. This surge has imposed operational difficulties on numerous hotels, necessitating the formulation of strategies to navigate these adverse circumstances (GSO, 2020). The significance of this matter within the Vietnamese context is profound. Tourism plays a pivotal role in Vietnam's economic landscape, contributing significantly to the nation's Gross Domestic Product (GDP). Addressing the challenge of booking cancellations is of paramount importance, as it offers numerous advantages for both travelers and hoteliers. Effectively addressing this issue fosters trust among visitors and ensures the sustainable growth of the tourism sector in the long term, particularly in Vietnam (Huy Lê, 2021), (Hà Phương, 2021).

A prominent Swedish institution has undertaken a comprehensive examination of the phenomenon of hotel reservation cancellations, resulting in a series of research papers dedicated to this inquiry. Enok Gartvall and Oscar Skanhagen's article from 2021 provides an extensive elucidation of the factors that precipitate hotel reservation cancellations (Gartvall

and Skanhagen, 2021). Their research is built upon the "Predicting Hotel Cancellations" dataset curated by Nuno Antonio, Ana Almeida, and Luis Nunes, and it deploys computational techniques such as Random Forest, XGBoost, and Logit models. These models, specifically Random Forest and XGBoost, are tree-based classifiers categorized as ensemble learning methodologies used to make qualitative predictions by iteratively conducting binary splits. Meanwhile, the logit model, or logistic regression, is employed as a benchmark model within this investigation. The primary findings of this research reveal that the Random Forest model outperforms the others when applied to hotel data, achieving an accuracy rate of nearly 80%. A significant determinant identified within the Random Forest model is the waiting time, a continuous variable representing the temporal gap between booking the hotel and the actual check-in date, which strongly influences the model's predictive performance. Furthermore, the inclusion of meteorological data has shown a modest improvement in predictive accuracy across all employed models for forecasting hotel reservation cancellations. Research from other nations, such as China, has also contributed valuable insights into the issue of hotel reservation cancellations. In a study authored by Yiying Chen, Chuhan Ding, Hanjie Ye, and Yuchen Zhou in 2022, the authors utilized the same "Predicting Hotel Cancellations" dataset and examined various algorithms, including logistic regression, k-nearest Neighbor classifier, and CatBoost (Chen et al., 2022). Their findings suggest that CatBoost is the most suitable predictive model for forecasting hotel reservation cancellations, as it consistently outperforms other models in terms of accurately predicted samples and overall accuracy. Similarly, Indonesia has contributed to this academic discourse with a study published in March 2021 (Putro et al., 2021). In this research, the authors explored the hotel booking cancellation issue using the "Hotel Booking Demand" dataset and applied Deep Neural Network and Logistic Regression algorithms. They found that by strategically removing influential attributes and employing the Logistic Regression algorithm, the predictive accuracy was enhanced from 79.66% to 80.29%. Furthermore, the discussion extends to the origin of the datasets used in these studies. The datasets were initially compiled in Portugal in 2017, sourced directly from hotel Property Management System (PMS) databases using Microsoft SQL Server (Nuno Antonio et al., 2017). Due to the consistent adoption of the same PMS application across all hotels, data uniformity was maintained. The study explored five distinct algorithms, with the Decision Forest algorithm, particularly when accompanied by precisely defined features, emerging as an effective strategy for crafting predictive models related to hotel booking cancellations. Finally, academic research in Thessaloniki, Greece, conducted in 2020, delved into the subject of hotel booking cancellations. The author sourced data from a 4-star hotel in Greece due to the limited availability of non-shared datasets (Timamopoulos, 2020). The study employed eight algorithms, including Logistic Regression, Naïve Bayes, K-Nearest Neighbors, Support Vector Machines (SVM), Decision Tree, Random Forests, Gradient Boosting Machines, and Extreme Gradient Boost (XGBoost). The findings emphasized that combining the SMOTE method with these algorithms yielded optimal performance for the dataset, achieving an accuracy rate of 99.93%.

The primary objective of this investigation is to conduct a comparative analysis of significant machine learning classifiers within the framework of evolving learning methodologies, contrasting them with conventional static machine learning approaches observed in prior research. This decision arises from the imperative need to embrace a continuous machine learning framework as opposed to a static model. This necessity arises due to the sustained utilization of the hotel cancellations dataset for future research endeavors pertaining to hotel cancellations in Vietnam. Another crucial rationale is associated with the concept of Concept Drift, a prominent phenomenon in the field of Machine Learning. Concept Drift refers to the temporal evolution of underlying conceptual structures. In the context of this study, it specifically refers to the dynamic changes in the conceptual foundation related to room cancellations over time. In predictive analytics, data science, machine learning, and related domains, Concept Drift represents the transformation of data that undermines the stability of a data model. This phenomenon occurs when the statistical characteristics of the target variable, which the model aims to predict, undergo unexpected changes over time, leading to a deterioration in prediction accuracy. Detecting and adapting to Concept Drift are essential in dynamic data environments. An illustrative example, often encountered in economic scenarios, relates to the changing consumer behavior in an online retail setting. For instance, when forecasting weekly merchandise sales, an initially effective predictive model may consider factors such as advertising expenditures, ongoing promotions, and other influential metrics. However, over time, the model's accuracy is likely to decline due to the manifestation of Concept Drift. In the context of merchandise sales, one factor contributing to Concept Drift could be seasonality, influencing variations in purchasing behavior with changing seasons, such as increased sales during winter holidays compared to summer months. Concept Drift arises as the variables comprising the dataset lose their explanatory power, possibly due to the emergence of new confounding variables, resulting in a gradual decline in accuracy. Recommendations suggest periodic assessments as part of post-production analysis and the retraining of models based on revised assumptions when Concept Drift indicators are identified. It is noteworthy that the majority of the aforementioned studies have refrained from incorporating Concept Drift, despite its necessity. This is in spite of it being a prerequisite for effectively applying a dataset in a context distinct from its original environment, as illustrated in this study. This course of action is attributed to the accumulation of new data, which necessitates immediate implementation due to the impracticality of waiting for data sufficiency to enable comprehensive machine learning.

This study is centered on an emerging research area that aligns with an innovative algorithmic framework. Its primary aim is to contribute to solving the challenge of hotel room cancellations. To achieve this objective, the study relies on a meticulously curated database and is focused on identifying a suitable method within this dataset to serve as the

foundational algorithm for a classification system. The ultimate goal is to provide organizations with the necessary tools, research insights, and practical outcomes to effectively address the issue of room cancellations through the utilization of an artificial intelligence system. This, in turn, will enhance the reliability of room reservations for both domestic and international establishments, not only in Vietnam but also in other regions. The research endeavor seeks to leverage artificial intelligence systems to predict customer room booking needs. This approach aims to improve businesses' understanding of tourist requirements, ultimately leading to increased revenue and fostering economic progress, both within Vietnam and on a global scale.

The research is structured as follows: The study commences by a state-of-art to provide insights into the concept of concept drift, a significant aspect of the study, it outlines the methodologies associated with concept drift that is important in the hotel cancellation problem; the research also examines the historical evolution of research related to room cancellations, with a specific focus on analyzing the datasets that have been employed in predictive efforts; the empirical segment of the research unfolds by implementing concept drift techniques across various algorithms for predicting hotel room cancellations to involves a comprehensive analysis and explanation of the results obtained through these techniques; the research introduces a demonstration system designed to familiarize researchers with the practical application of the developed algorithms and systems; Finally, the research concludes, summarizing the key findings and highlighting the significance of the study's contributions to the field of hotel room cancellation prediction. It also emphasizes the potential impact of the research on the hospitality industry and the broader economy. In essence, this study aims to pave the way for more accurate and efficient room reservation systems, benefiting both businesses and travelers, not only in Vietnam but also in various international contexts.

2. STATE-OF-ART OF EVOLVING MACHINE LEARNING IN ARTIFICIAL INTELLIGENCE

In the contemporary landscape, traditional databases grapple with challenges arising from their inherent inflexibility over time. This limitation primarily stems from their reliance on classical algorithms, which are characterized by single-run executions. These algorithms mandate a complete retraining process when new data becomes available. For instance, consider a scenario in which data 1 is utilized for model training. Upon the arrival of new data 2, both data 1 and data 2 require retraining in order to construct a fresh model. However, this conventional approach encounters constraints in modern dynamic settings, where data ecosystems are in a constant state of evolution. In such contexts, the ability to adapt in real-time and update models becomes indispensable, highlighting the necessity for continuous learning within environments marked by perpetually changing data. This conceptual framework is referred to as "continuous learning."

Various methodologies have been employed to extend classical algorithms into the realm of continuous learning paradigms, often achieved through the integration of sliding window techniques. A comprehensive exploration of these methodologies has been extensively addressed in the scholarly work authored by Ngo Ho (Ngo Ho, 2013). As a result, the pursuit of evolving concepts finds a suitable embodiment through the application of a straightforward "sliding windows" approach, reminiscent of the FLORA approach pioneered by Widmer et al. (Widmer et al., 1996). The fundamental principle involves the periodic updating of the model at each discrete moment 't' using the most current training data, demarcated by a sliding window of specified dimensions, whether it be a temporal scale or a specific count of data points. This methodology can accommodate either batch retraining, utilizing the data contained within the sliding window, or model updates, depending on compatibility with online learning methods. This paradigm typically consists of a sequence of three steps, as proposed by Bifet et al. (Bifet et al., 2007):

1. Detection of concept alterations through the application of statistical tests conducted on distinct windows.
2. In cases where changes are observed, the curation of representative and contemporaneous exemplars is performed to facilitate model adaptation.
3. Subsequent model updates are carried out to embody the evolving understanding of the data distribution.

The dimensions of the window are predetermined by the user in advance, with each successive window having partial overlap with its predecessor, thereby sharing a portion of the data. During each iteration, a new model is obtained, which inherently represents an updated class ensemble. The core essence of these methodologies hinges on the careful choice of the window size. While many strategies utilize a fixed-size window tailored to each distinct real-world scenario, there are methodologies focused on automating the determination of the window's dimensions.

Several approaches have been introduced in the literature for managing and adapting window sizes in data stream processing, each with its unique characteristics and advantages. Here, we briefly describe some of these methods:

1. **ADWIN (ADaptive WINdow)**: Bifet and his colleagues introduced ADWIN, which involves testing an array of window sizes by dividing each window into smaller sub-windows of a minimum yet meaningful size. If these sub-windows exhibit significantly different distributions, the specific size is considered an appropriate choice based on rigorous statistical evaluation (Bifet et al., 2007).
2. **Bifurcated Approach**: Lazarescu and his team advocate a bifurcated approach where two models are maintained at each iteration. These models use different window sizes: one with a standard size denoted as S and the other

with an augmented size of 2S. The standard-sized window is used to detect new conceptual shifts through statistical assessments, while the larger window (2S) is employed to update the model when novel conceptual transformations are detected (Lazarescu et al., 2004).

3. **OLIN (On Line Information Network) Methodology**: Presented by Last and colleagues, the OLIN methodology is characterized by dynamically adjusting the window size based on performance feedback from a validation dataset. At each stage, the newly acquired data is divided into training and validation segments. An array of windows with varying sizes is independently applied to each segment, and the size that yields the best results on the validation dataset is selected for the next stage. It's worth noting that this approach requires conducting learning phases on sufficiently large data batches (Last et al., 2002).
4. **Incremental Window Sizes**: Klinkenberg and his team propose an incremental progression of window sizes. They evaluate performance (usually in terms of error rate) across various window extents, such as size *No1* representing the latest batch, size *No2* for the last two batches, size *No3* for the last three batches, and so on. The window size that provides the best performance is adopted for subsequent iterations (Klinkenberg et al., 2004).

These methods provide diverse strategies for adapting window sizes in data stream processing, catering to different needs and scenarios, such as detecting concept drift or optimizing performance. The choice of method depends on the specific requirements and characteristics of the data stream application at hand.

In this discussion, we embark on an exploration of a diverse set of six distinct machine learning algorithms, enhanced by the incorporation of progressive strategies inspired by the Klinkenberg concept, to dynamically determine the optimal window size. Here, we provide explanations of these machine learning algorithms:

1. **Gaussian Naïve Bayes** (Chan et al., 1979): A variant of the Naïve Bayes classifier that assumes features follow a Gaussian (normal) distribution. This algorithm is particularly suitable for continuous data and has the capability to continuously update feature means and variances as new data arrives.
2. **Decision Tree Classifier** (Breiman et al., 1984): Decision trees are structured like flowcharts, where internal nodes represent features or attributes, branches represent decision rules, and leaf nodes contain outcome predictions. The classifier learns how to partition data based on the values of these attributes.
3. **K-Neighbors Classifier** (Goldberger et al, 2005): A method that finds a predefined number of training samples closest in distance to a new point and predicts the label based on these samples. The number of samples can be specified by the user..
4. **Linear Discriminant Analysis** (McLachlan, 2004): Linear Discriminant Analysis (LDA) aims to discover a linear combination of distinct features that effectively differentiate between two or more categories or classes of entities or events. The resulting combination can serve as a dimensionality reduction technique before subsequent classification tasks.
5. **Bernoulli Naive Bayes** (Lewis, 1998): Bernoulli Naive Bayes is a specialized variation of the Naïve Bayes classifier, designed for binary data representations. It is particularly useful when dealing with data that has been transformed into binary form, especially in certain situations.
6. **Multi-layer Perceptron Classifier** (Hinton et al, 1989): Relies on a neural network to perform classification tasks, optimizing the log-loss function using LBFGS or stochastic gradient descent.

These machine learning algorithms, coupled with adaptive strategies inspired by the Klinkenberg concept, form a robust toolkit for dynamic window size adaptation in the context of data streams with evolving characteristics and concept drift. This approach aims to enhance the adaptability and efficiency of the algorithms when processing streaming data.

The core of this investigation hinges on the fusion of these learning methodologies with the "*sliding windows*" approach, effectively implementing Klinkenberg's innovative concept for real-time determination of the optimal window size. The algorithms utilized in this experimental study were instantiated using version 0.24.2 of the scikit-learn library, integrated with the "*sliding windows*" methodology. The dataset undergoes a progressive, stream-based processing approach, employing small instances of "*sliding windows*" that encapsulate the most recent n samples. The experimentation begins with n set at 1, representing the classical online learning paradigm. Subsequently, the value of n is systematically increased beyond 1, indicating a transition toward batch learning scenarios. For each distinct method, only the most optimal results corresponding to various values of n are selected for subsequent comparison with alternative learning techniques.

3. ANALYSIS OF DATASETS AND PERFORMANCE INDICATORS

During the data acquisition process within the research domain, a multitude of datasets were encountered, but only three exhibited noteworthy attributes, distinguished by their comprehensive variables and abundant availability. These selected datasets were earmarked for investigation in the context of research on hotel reservation cancellations. They include the following:

- "Reservation Cancellation Prediction" Dataset (Gaurav Dutta, 2022): This dataset, authored by Gaurav Dutta, comprises approximately 36,200 rows containing 18 fields. Gaurav Dutta was responsible for both collecting and synthesizing this dataset. Published in December 2022 and licensed under Attribution 4.0 International (CC BY 4.0), it includes two distinct files: the train_dataset with an approximate row count of 18,100, and the test_dataset, which features a comparable number of rows.
- "Hotel Booking Demand" Dataset (Jesse Mostipak, 2020): This dataset, conceived by Jesse Mostipak, is derived from the original "Predicting Hotel Cancellations" dataset. However, this derivative possesses fewer attributes than its progenitor. It encompasses 32 fields and incorporates 119,000 data entries. Published in 2020 and licensed under Attribution 4.0 International (CC BY 4.0), its utility may be constrained by its temporal origins, potentially rendering it outdated relative to the contemporary context.
- "Hotel Booking" Dataset (Mojtaba, 2021): The dataset "Hotel Booking," authored by Mojtaba, includes 36 fields and approximately 119,000 rows. Mojtaba synthesized and curated this dataset, drawing from the "Predicting Hotel Cancellations" dataset. Published in 2021, it is licensed under Data files © Original Authors. Despite its 2021 release date, this dataset may be considered relatively antiquated, potentially impacting its suitability for experimental investigations.
- "Hotel Reservations Dataset" (Ahsan Raza, 2023): Authored by Ahsan Raza, the "Hotel Reservations Dataset" emerged on February 12, 2023. This dataset comprises 36,275 data entries distributed across 19 fields and operates under the Attribution 4.0 International (CC BY 4.0) license.

Among the four datasets considered, only one proves to be suitable for classification due to data availability and contemporary relevance. The "Hotel Reservations Dataset" authored by Ahsan Raza, most recently updated on February 12, 2023, has been chosen as the primary dataset for integration into the research framework. The selection is motivated by the dataset's up-to-date attributes, making it a pragmatic choice for classification endeavors. This dataset encompasses a variety of features, including:

- **Label:** This feature represents predicted results, where 0 stands for no booking cancellation, and 1 stands for booking cancellation.
- **no_of_adults**: Indicates the number of adults who have booked a room, with a minimum value of 0 and a maximum of 4.
- **no_of_children**: Represents the number of children included in the booking, with a minimum of 0 and a maximum of 10.
- **no_of_weekend_nights**: Specifies the number of weekend nights (Saturday or Sunday) the guests stayed or booked at the hotel, with a minimum of 0 and a maximum of 7.
- **no_of_week_nights**: Represents the number of weekday nights (Monday to Friday) the guests stayed or booked at the hotel, with a minimum of 0 and a maximum of 17.
- **type_of_meal_plan**: Indicates the type of meal plan the customer has booked to use during their stay, ranging from "Not Selected" to "Meal Plan 3."
- **required_car_parking_space**: This feature indicates whether the customer requested a parking space during their stay, with values of 0 (no) or 1 (yes).
- **room_type_reserved**: Specifies the type of room the guest has booked for their stay, with values ranging from "Room_Type 1" to "Room_Type 7."
- **lead_time**: This feature represents the time from the booking date to the guest's arrival date at the hotel, with a minimum of 0 and a maximum of 443.
- **arrival_year**: Indicates the year the customer arrived at the hotel, with values ranging from 2017 to 2018.
- **arrival_month**: Represents the month the customer arrived at the hotel, with values ranging from 1 to 12.
- **arrival_date**: Indicates the day of the month the customer arrived at the hotel, with values ranging from 1 to 31.
- **market_segment_type**: Specifies the market segment, including options like Online, Offline, Corporate, Complementary, and Aviation.
- **repeated_guest**: Indicates whether the guest has stayed at the hotel before, with values of 0 (no) or 1 (yes).
- **no_of_previous_cancellations**: Represents the number of previous bookings that the customer had canceled before the current booking, with a minimum of 0 and a maximum of 13.
- **no_of_previous_bookings_not_canceled**: Indicates the number of previous bookings that the customer did not cancel before the current booking, with a minimum of 0 and a maximum of 58.
- **avg_price_per_room**: Represents the average price per room per day, with room prices fluctuating in euros, and values ranging from 0 to 540.
- **no_of_special_requests**: Indicates the number of special requests made by the guest when booking the room at the hotel, with values ranging from 0 to 5.

This dataset aligns well with the research criteria, requiring minimal modifications to the data. It will be stored in .csv file format, and after applying artificial intelligence techniques, it will yield a model file with a .sav extension for further

analysis and experimentation.

The evaluation of results universally relies on the metric of accuracy. This assessment method is characterized by its simplicity, as it quantifies the ratio of correct predictions to the total number of data points within the test dataset. Accuracy is typically expressed as a percentage, calculated by dividing the number of correctly predicted outcomes by the total dataset size. While this approach is initially effective, it can encounter challenges when dealing with severe data imbalances, often referred to as imbalanced datasets (phamdinhkhanh, 2020). For example, scenarios involving skewed class ratios, such as a 90:10 distribution or a situation where 100 individuals are involved, with 99 having a particular condition while only one does not, can lead to a misinterpretation of the model's effectiveness. In such cases, the accuracy metric may yield a high value, which may not provide meaningful insights into the model's performance.

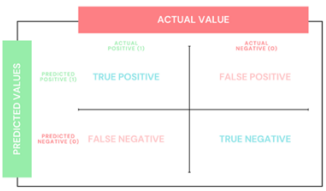

Fig 1 Confusion Matrix

In the matrix above, the terms "positive" or "negative" in TP/FP/TN/FN refer to the predictions made, not the actual labels. (Hence, a "false positive" indicates an incorrect positive prediction.) Here are the formulas for sensitivity and specificity based on the confusion matrix:

Sensitivity formula:

Sensitivity= TP / (TP + FN)

Specificity formula:

Specificity =TN / (TN + FP)

Therefore, Balanced Accuracy was introduced to address the aforementioned imbalance scenario. To resolve the issue in the given example, a calculation method was devised based on the Confusion Matrix, from which True Negative Rate (TNR), True Positive Rate (TPR), False Negative Rate (FNR), and False Positive Rate (FPR) can be calculated. After obtaining all these metrics, the formula of Balanced Accuracy can be used to compute the most realistic and optimized percentage.

Balanced Accuracy formula:

$$\textbf{Balanced accuracy} = \frac{(\textbf{Sensitivity} + \textbf{Specificity})}{2}$$

4. RESULTS AND DISCUSSION

The experimental setup you've described involves the use of a training dataset and a testing dataset for evaluating the performance of a classification model. Here are the key details of the implementation method:

Training and Testing Data

Training Dataset: This dataset consists of 25,393 instances, which constitutes 70% of the original data. It is used to train the classification model.

Testing Dataset: The testing dataset contains 10,882 instances, constituting 30% of the original data. This dataset is used to evaluate the model's performance.

Random Shuffling

Data instances are shuffled randomly both before and after the training process. This randomization helps ensure that the model is not biased by the order of data instances and can generalize well to unseen data.

Indirect Experimental Model (Batch Learning)

The experiment is conducted using an indirect experimental model with a batch size of 12,695. This means that the system will execute 12,695 steps, and each step involves processing approximately one data instance.

The batch size chosen is equivalent to 70% of the original data, which is the same as the training dataset's size. This choice can enhance accuracy, particularly with large and complex datasets.

Processing Time Consideration

Increasing the batch size to 70% of the original data size can indeed improve accuracy, but it comes at the cost of requiring a substantial amount of processing time. Larger batch sizes require more computational resources and time for training.

Discussion

Overall, this experimental setup aims to achieve a balance between accuracy and computational efficiency by using a sizable training dataset and an indirect experimental model with batch learning. The random shuffling of data instances helps ensure the model's robustness and ability to handle diverse data patterns.

The dataset employed in the experiment comprises two distinct segments: a training set and a test set. The training set encompasses 25,393 data rows, while the test set comprises 10,882 data rows. It is important to note that while the dataset may change in each experiment, the data remains unchanged, with only shuffling applied (data preservation). The model utilized in the experiment is a batch data model. This model operates by partitioning the initial dataset into smaller batches, each containing 12,695 data instances. This batching approach is chosen for its manageability and efficiency in terms of time consumption, ensuring optimal outcomes for the problem at hand.

In the following, we apply specific artificial intelligence methods, namely algorithms such as MLPClassifier, KneighborsClassifier, Linear Discriminant Analysis, BernoulliNB Classifier, Decision Tree Classifier and GaussianNB Classifier, in conjunction with the sliding window technique. The experimental results of these algorithms are presented through diagrams below:

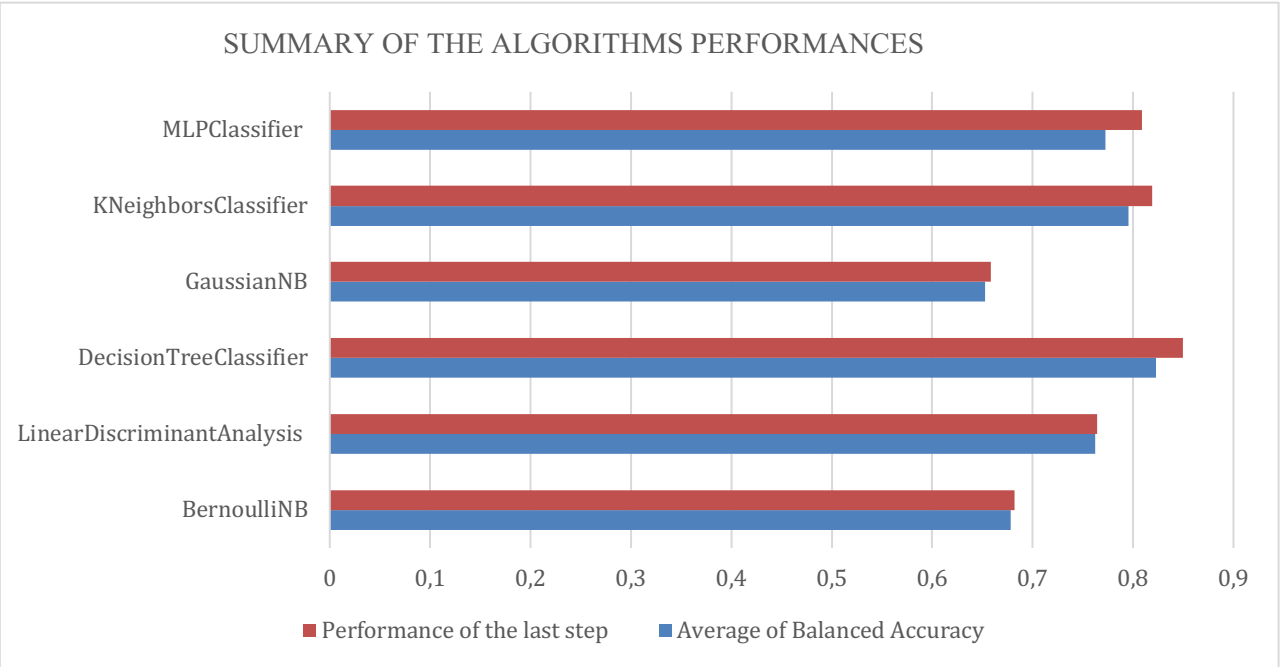

Fig 2 Summary of the algorithms performances

Examining the chart, it is evident that when considering the final step and the overall average across all six algorithms, the DecisionTreeClassifier algorithm performs the best in the hotel room reservation prediction task compared to the other five algorithms. The remaining five algorithms are LinearDiscriminantAnalysis, BernoulliNB, MLPClassifier, KneighborsClassifier, and GaussianNB. When comparing them to each other, it can be observed that both GaussianNB and BernoulliNB algorithms have a final step and an average around below 70%. Specifically, GaussianNB reaches both the final step and the average at around 65%, while BernoulliNB has a final step of 68% and an average of 67%. The next algorithm, LinearDiscriminantAnalysis, has a final step and an average below 80%, but not significantly so. For LinearDiscriminantAnalysis, both the final step and the average reach 76%. Moving on to the MLPClassifier algorithm, despite having a final step of 80%, the average is not as high, only reaching 77%. The remaining algorithm, when compared to the DecisionTreeClassifier, shows differences. Specifically, the KneighborsClassifier has a final step of 81% and an average of 79%, while the DecisionTreeClassifier has a final step of 84% and an average of 82%.

Comparing the six algorithms to each other, we can observe that both GaussianNB and BernoulliNB algorithms have the lowest final step and average compared to the other four algorithms. Next, the LinearDiscriminantAnalysis algorithm has a higher final step and average compared to the two previous algorithms but is still lower than the other three algorithms. Following that, the MLPClassifier algorithm also has a higher final step and average than the three previous algorithms but not significantly higher than the LinearDiscriminantAnalysis algorithm. Finally, the KneighborsClassifier and DecisionTreeClassifier algorithms, when compared to the previous four algorithms, are higher. However, when looking

at the chart, it is evident that only the DecisionTreeClassifier algorithm stands out as significantly higher than the other four algorithms, while the KneighborsClassifier does not achieve the same level of distinction. Specifically, the KneighborsClassifier algorithm is notably higher than the three algorithms: GaussianNB, BernoulliNB, and LinearDiscriminantAnalysis. However, concerning the MLPClassifier algorithm, there is not a significant difference. In conclusion, when considering the overall performance, the DecisionTreeClassifier algorithm appears to be the best-suited algorithm for practical application.

In addition to calculating the average results of the algorithms, another approach is to compare the experimental model results by age group, which provides a more comprehensive and detailed perspective. This approach helps us visually assess and arrive at the most accurate conclusions. The experimental model results by age group are presented in the chart below:

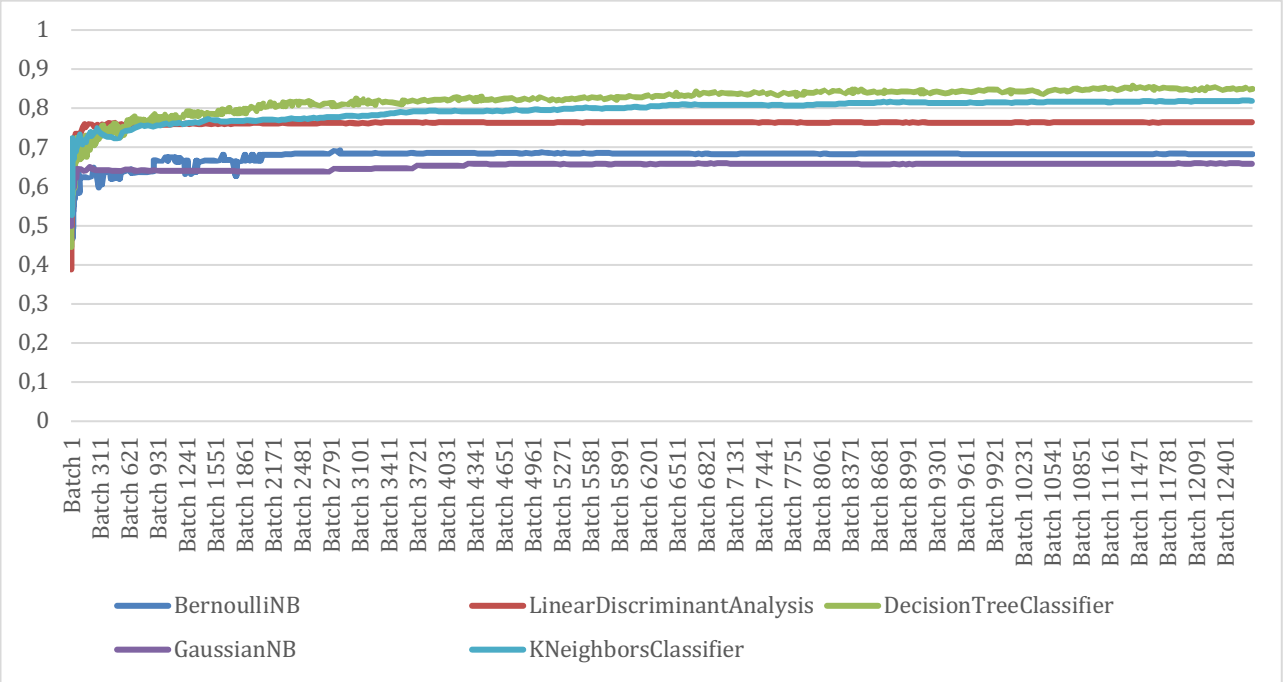

Fig 3 Learning progression of the algorithms

Looking at the overall chart and considering the variability, it's evident that the DecisionTreeClassifier algorithm has the lowest starting point, beginning at around 46%. However, it gradually stabilizes and reaches a maximum of 84.9% at the last batch, which is the highest achieved by the DecisionTreeClassifier algorithm. Next is the GaussianNB algorithm, which starts slightly higher, around 49%, but exhibits instability during its ascent. However, it stabilizes at batch 4302 and reaches a maximum of 65%. Following that, both the BernoulliNB and KneighborsClassifier algorithms start at 66%, but they have different patterns of progress. The KneighborsClassifier algorithm steadily and slowly rises and eventually reaches a maximum of 81% and maintains this level until the end. In contrast, the BernoulliNB algorithm also climbs slowly but lacks the stability seen in the KneighborsClassifier, reaching a maximum of 68% at batch 7096 and maintaining it until the end. The MLPClassifier algorithm starts at 67% and reaches a maximum of 80% quite early, at batch 289. However, its variability is noticeable as it fluctuates throughout the experiment, indicating a degree of instability. Regarding the LinearDiscriminantAnalysis algorithm, it also starts at 67% and reaches a maximum of 76% at batch 3155. While it achieves a lower maximum compared to the MLPClassifier, it maintains stability throughout the experiment. Despite the DecisionTreeClassifier algorithm having the lowest starting point, it takes considerable time to reach its maximum. Although it starts relatively low, it surpasses the other five algorithms from batch 3088 onwards. In conclusion, considering the variability of all six algorithms, the GaussianNB algorithm starts relatively low but reaches a maximum similar to BernoulliNB. LinearDiscriminantAnalysis starts relatively high but ends at 76%. The MLPClassifier, while starting high and reaching its maximum early, exhibits noticeable instability. The KneighborsClassifier starts at the same level as BernoulliNB but surpasses it and the other four algorithms, reaching 81%. Finally, the DecisionTreeClassifier, despite starting with the lowest value, outperforms the others, especially from batch 3088 onwards.

In summary, the DecisionTreeClassifier algorithm appears to be highly suitable for the hotel room reservation prediction problem and practical applications.

5. IMPLEMENTATION

With the introduction of this application, significant benefits are expected to accrue, contributing to a substantial

improvement in the economic situation of hotel businesses in Vietnam. Building upon the conclusive outcomes outlined in the previous section, the research team proceeded to integrate a version of the K-Nearest Neighbors (KNN) algorithm into the predictive system. This system is equipped with functional nodes encompassing prediction capabilities, execution of classical algorithms, a repository for processed models, system configuration options, and user authentication features. It is designed for deployment within a web-based environment, offering two principal functionalities: algorithm installation, accessible to administrators or developers, and diagnostics, accessible to users. These functionalities are further detailed in the subsequent use case diagram depicted as Figure 4.

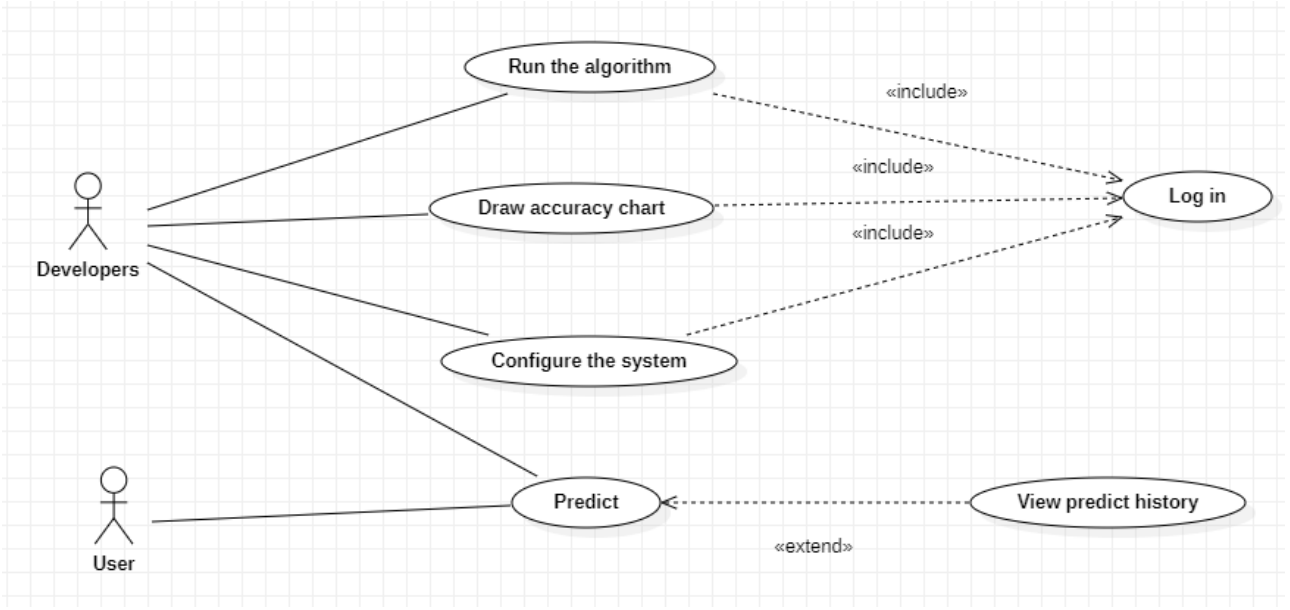

Fig 4 System Use Case Diagram

Regarding the installation of the system on a computing platform, the installation process involves the download of the designated installation file, labeled as "HotelReservationsDataset.rar." Upon extraction, a directory named "HotelReservationsDataset" will be created. To run the software, the user's computer must have specific Python libraries installed, along with Python version 3.9.9. After extraction, within the "SETUP" directory, you will find a file named "python-3.9.9-amd64.exe," which is used for installing Python 3.9.9, and "inLib.bat," which is used to install the necessary libraries for software execution. The environment setup is completed with the inclusion of a "Remove.bat" file, which is used to delete extraneous data files, including test-run records. This file should only be executed in two instances: immediately after extraction and installation, or when there is a need to erase all previously executed data. To initiate the execution of the program, the "Runserver.bat" file is used. This file is configured to trigger the command "py manage.py runserver," thus starting the program on the default port "http://127.0.0.1:8000/." It's important to note that the user's system must have a stable internet connection and meet certain minimum system requirements, including Windows 10, 2GB of RAM, and a hard drive with a capacity exceeding 10GB, to ensure smooth operation. Following a successful installation, the software can be accessed by navigating to the designated port "http://127.0.0.1:8000," providing access to the main interface of the system. This interface prominently features multiple input fields designed for predictive purposes. Below is an illustration of the primary interface of the "Hotel Reservation Prediction" system, accessible at http://hrp.adhigtechn.com/.

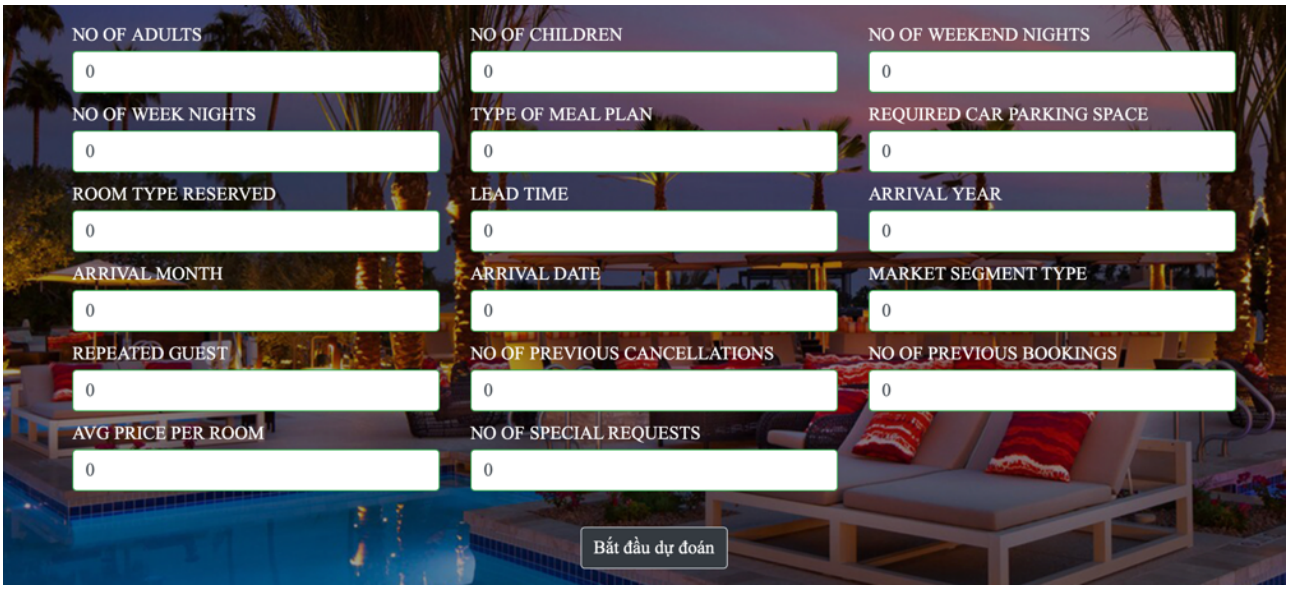

Fig 5 Main interface of the prediction system (demo*)*

5. CONCLUSION

Upon the completion of the research process and the compilation of the report, a comprehensive evaluation of the obtained results becomes feasible. The report's content is characterized by clarity, providing specific explanations of the data, charts, and algorithms. In particular, the system has successfully implemented the classical KNeighborsClassifier algorithm, addressing the challenges of training models, making predictions, and handling variable data in dynamic environments.

This topic is poised for further development in the future, which may involve updating the dataset through survey methods and incorporating specialized knowledge to acquire the most standardized and closely aligned dataset with reality. The system's functionalities have been thoroughly implemented, effectively meeting the initial requirements. It is designed with separate sections for general users and developers, offering a user-friendly interface for predicting hotel reservations while providing developers with additional pages for system enhancement and development.

6. ACKNOWLEDGMENTS

The authors gratefully acknowledge the contribution of Dr. Huynh Thanh NGUYEN, Director of Institute of Developmental Philosophy for his encouragement and extraordinary support throughout this research.

7. REFERENCES

A.Bifet, and R. Gavalda (2007). Learning From Time-Changing Data With Adaptive Windowing ; In: Sdm, Pp. 443–448.

Chan, Golub, and LeVeque (1979). Updating formulae and a pairwise algorithm for computing sample variances, Stanford CS tech report STAN-CS-79-773, Stanford University.

Christos Timamopoulos (2020). Anomaly Detection: Predicting hotel booking cancellations. A thesis submitted for the degree of Master of Science (MSc) in Data Science, School of Science & Technology, Thessaloniki, Greece (January 2020).

D. Bjorkegren and D. Grissen (2015). Behavior revealed in mobile phone usage predicts loan repayment.

David D. Lewis (1998). Naive (Bayes) at forty: The independence assumption in information retrieval, Machine Learning: ECML-98, Lecture Notes in Computer Science book series (LNAI,volume 1398).

Enok Gartvall, Oscar Skanhagen (2021). Predicting hotel cancellations using machine learning. Journal of Travel Research, Bachelor Thesis in Statistics, School of Business, Economics and Law, Department of Economics, September 2021.

G.Widmer, M.Kubat (1996). Learning In The Presence Of Concept Drift And Hidden Contexts ; Machinelearning, Vol. 23, No. 1, Pp. 69-101.

GSO (2022). Socio-Economic Situation in The First Quarter of 2020, General Statistics Office Of Vietnam, Date of Issue: 31/03/2020.

Hà Phương (2021). Vai trò của du lịch trong Chiến lược phát triển kinh tế - xã hội của Thủ đô Hà Nội, Tạp chí cộng sản

(ISSN 2734-9071), Quốc Phòng - An Ninh - Đối Ngoại, 25-10-2021.

Hinton, Geoffrey E (1989). Connectionist learning procedures. Artificial intelligence 40.1 (1989): 185-234.

Huy Lê (2021). Du lịch Việt Nam: Nỗ lực chuyển mình, chủ động thích ứng trong tình hình mới, Báo điện tử Đảng cộng sản Việt Nam. Retrieved: 21/08/2023. Original: 09/07/2021. Link: https://dangcongsan.vn/kinh-te/du-lich-viet-nam-no-luc-chuyen-minh-chu-dong-thich-ung-trong-tinh-hinh-moi-584986.html.

J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov (2005). Neighbourhood Components Analysis, Advances in Neural Information Processing Systems, Vol. 17, May 2005, pp. 513-520

J. Kleinberg, J. Ludwig, S. Mullainathan, and Z. Obermeyer (2015). Prediction policy problems. American Economic Review, 105(5):491–95, 2015.

L. Breiman (1996). Bagging predictors, Machine Learning, vol. 24(2), pp.123-140.

M.Last (2002). Online Classification Of Nonstationary Data Streams ; Intell. Data Anal. Vol.6(2), pp.129–147.

M.Lazarescu, S.Venkatesh, and H.Bui (2003).Using Multiple Windows To Track Concept Drift ; vol.Ida8(1), Pp.29–59.

McLachlan, G. J. (2004). Discriminant Analysis and Statistical Pattern Recognition. Wiley Interscience. ISBN 978-0-471-69115-0. MR 1190469

Ngo-Ho Anh-Khoi (2015), Méthodes de classifications dynamiques et incrémentales : application à la numérisation cognitive d'images de documents, Doctoral dissertation, Ecole doctorale Mathématiques, Informatique, Physique Théorique et Ingénierie des Systèmes (Centre-Val de Loire), Tours, 2015.

Nugroho Adi Putro, Rendi Septian, Widiastuti, Mawadatul Maulidah, and Hilman Ferdinandus Pardede (2021). Prediction of Hotel Booking Cancellation Using Deep Neural Network And Logistic Regression Algorithm. STMIK Nusa Mandiri. Jurnal Techno Nusa Mandiri Vol.18, No.1.

Nuno Antonio, Ana Almeida, Luis Nunes (2017). Predicting Hotel Booking Cancellations To Decrease Uncertainty And Increase Revenue, Tourism & Management Studies (ISSN: 2182-8458), vol. 13, núm. 2, Universidade do Algarve Faro, Portugal, 2017, pp. 25-39

R.Klinkenberg (2004). Learning Drifting Concepts: Example Selection Vs. Example Weighting ; Intelligent Data Analysis, Special Issue On Incremental Learning Systems Capable Of Dealing With Concept Drift, vol.8 (3).

Susan Athey (2018). The Impact of Machine Learning on Economics, University of Stanford.

Tapscott and C. Don (1996). “The digital economy: promise and peril in the age of networked intelligence,” Educom Review, vol. 12, 1996.

Yiying Chen, Chuhan Ding, Hanjie Ye, and Yuchen Zhou (2022). Comparison and Analysis of Machine Learning Models to Predict Hotel Booking Cancellation, Advances in Economics, Business and Management Research (ISSN 2352-5428), Volume 211, DOI 10.2991/aebmr.k.220307.225.