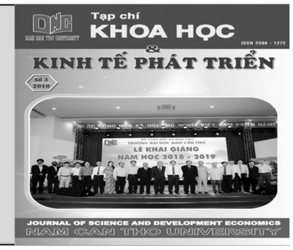
Journal of Science and Development Economics
Nam Can Tho University

Website: jsde.nctu.edu.vn



Performance of milk quality diagnostics using extra tree classifier techniques with progressive learning

Pham Hoang Minh[1], Truong Hung Chen[1], Pham Huynh Thuy An[2], Ngo Ho Anh Khoi[1]
[1]*Faculty of Information Technology, Nam Can Tho University*
[2]*Faculty of Engineering and Technology, Nam Can Tho University*

Received: November 10, 2023
Revised: November 30, 2023
Accepted: December 5, 2023

Keywords: *AI, diagnosis, extra tree classifier, milk quality, machine learning prediction*

Từ khoá: *bộ phân loại cây bổ sung, chẩn đoán, chất lượng sữa, dự đoán, học máy, trí tuệ nhân tạo*

ABSTRACT

Quality control of milk involves the use of established control measures and testing methods to ensure proper adherence to standards and regulations concerning milk and its products. Testing ensures that dairy products meet the requirements of standards, are acceptable in terms of nutritional content, and adhere to safety standards regarding microbiological factors, heavy metals, pesticide residues, veterinary drug residues, toxins, and more. Therefore, quality checks at various stages of the milk processing chain, from farms to processing facilities and consumers, are crucial. The research method involved scientific experimentation, conducted using the Extra Tree Classifier algorithm with evolving method. The scope of the study was not extensive, and the dataset was the Milk Quality Prediction dataset sourced from kaggle.com. The aim of the study was to aid in diagnosing milk quality rapidly and relatively reliably through provided numerical data. This endeavor aims to reduce the prevalence of low-quality milk trading, ultimately contributing to safer and higher quality milk management for consumers.

TÓM TẮT

Kiểm soát chất lượng của sữa liên quan đến việc sử dụng các biện pháp kiểm soát và phương pháp kiểm tra được thiết lập để đảm bảo tuân thủ đúng các tiêu chuẩn và quy định liên quan đến sữa và các sản phẩm của nó. Thử nghiệm đảm bảo rằng các sản phẩm sữa đáp ứng các yêu cầu về tiêu chuẩn, có thể chấp nhận được về hàm lượng dinh dưỡng và tuân thủ các tiêu chuẩn an toàn liên quan đến các yếu tố vi sinh, kim loại nặng, dư lượng thuốc trừ sâu, dư lượng thuốc thú y, độc tố,… Do đó, kiểm tra *chất lượng ở các giai đoạn khác nhau của chuỗi chế biến sữa, từ các trang trại đến các cơ sở chế biến và người tiêu dùng là rất quan trọng. Phương pháp nghiên cứu này bao gồm thử nghiệm khoa học, được thực hiện bằng cách sử dụng thuật toán phân loại cây thêm với phương pháp phát triển. Phạm vi của nghiên cứu không rộng và bộ dữ liệu được sử dụng là bộ dữ liệu dự đoán chất lượng sữa có nguồn gốc từ Kaggle.com. Mục đích của nghiên cứu là hỗ trợ chẩn đoán chất lượng sữa nhanh chóng và tương đối đáng tin cậy thông qua dữ liệu được cung cấp. Nỗ lực này nhằm mục đích giảm tỷ lệ giao dịch sữa chất lượng thấp, cuối cùng góp phần vào việc quản lý sữa an toàn và chất lượng cao hơn cho người tiêu dùng.*

1. INTRODUCTION

Milk quality is a crucial factor in milk production to ensure the safety of dairy products and their suitability for various purposes. To achieve quality, hygiene requirements must be implemented throughout the entire milk processing chain. Currently, with the global and specifically Vietnamese technological advancements, the counterfeiting of essential products has become an intricate issue. Dairy products, including milk, are not exceptions to this trend. Such counterfeiting can have detrimental effects on consumer health. Managing and assuring milk quality before it reaches consumers' hands is an immensely important and necessary matter, particularly in Vietnam. Milk is an essential commodity, especially for the elderly and children. It plays a vital role in enhancing health and supporting human development. Hence, if counterfeited, it can significantly impact well-being. Therefore, diagnosing and assessing milk quality hold great significance.

Elwell (2006) [1] discussed the method of using microfiltration to enhance milk quality. In this study, various milk parameters were highlighted as essential factors determining quality and extending the shelf life of milk. Marth (1981) [2], presented the interconnectedness of milk production and supply, emphasizing that ensuring high-quality milk and its derived products requires collaborative efforts from various stakeholders. The process involves producers, processors, distributors, and ultimately consumers. In another article, Marth (1972) [3] discussed the standards for quality testing of dairy products, shedding light on the criteria used to evaluate milk quality. Tường Vi (2019) [4] presentd stringent regulations to ensure milk quality before it reaches consumers' hands. In VIRAC (2017) [5], the author provided an overview of quality and food safety management, underscoring the importance of ensuring milk quality during the milk processing.

Additionally, many articles about the testing milk problem has been proposed. In Viện Kiểm Nghiệm An Toàn Vệ Sinh Thực Phẩm Quốc Gia (2022) [6], the author discussed the necessity of milk and dairy product testing, outlines regulations related to such testing, and explores the process of testing milk and dairy products at the National Institute for Food Safety Testing. Many of them works with the AI methods. Padmaja (2021) [7] utilized the Extra Tree Classifier algorithm along with various other algorithms. Specifically, the article proposed the use of machine learning classification algorithms to predict Vitamin D Deficiency (VDD) severity. These machine learning algorithms included Random Forest (RF), Multi-Layer Perceptron (MLP), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), Gradient Boosting (GB), Stochastic Gradient Descent (SGD), AdaBoost (AB), Extra Tree Classifier (ET), and Logistic Regression (LR). The article evaluated the output of different machine learning classification methods to estimate the severity of VDD in individuals. In the first two decades of the 21st century, a series of analyses, including Free Volatile Carboxylic Acids (FVCAs), attempted to describe 10 different types of cheese from Switzerland. The current work aimed to investigate whether these 10 types of cheese can be classified using supervised machine learning techniques and to analyze the significance of FVCAs features in understanding the role they play in describing the various types of cheese. Particularly, emphasis was placed on the SHAP (Shapley Additive exPlanations) values. In total, 241 cheese samples were classified using various ML algorithms with the assistance of the PyCaret library, with at least 90% accurately classified by two ensemble algorithms: Extra Tree and Random Forest (Bachmann, 2023) [8]. Mujumdara (2019) [9] compared multiple algorithms including Decision Tree, Gaussian NB, LDA, SVC, Random Forest, ExtraTrees, AdaBoost, Perceptron, Logistic Regression, Gradient Boost Classifier, Bagging, and KNN applied to the Diabetes Prediction dataset. This dataset consists of 800 rows and 10 fields: Number of Pregnancies, Glucose Level, Blood Pressure, Skin Thickness (mm), Insulin, BMI, Age, JobType (Office-work/Fieldwork/Machine-work), Outcome. The process includes several steps: (1) import necessary libraries, (2) load the dataset, (3) adjust paths in the training dataset, (4) compare accuracy, and (5) predict and determine the most accurate model on the test data. The employed approach can serve as a reference regarding the operational aspects of problem-solving.

2. MATERIALS AND METHODS

During the data search process for the topic, there were not a considerable number of datasets related to milk quality (approximately 135 relevant datasets). However, among them, three datasets with complete and high-quality information stood out: First, "Milk Quality Prediction" dataset by Rajendran (2022): This dataset comprises 1059 rows and 7 fields, including pH (ranging from 3 to 9.5), Temperature (ranging from 34°C to 90°C), Taste (with only 2 values: 0 and 1), Odor (with only 2 values: 0 and 1), Fat (with only 2 values: 0 and 1), Turbidity (with only 2 values: 0 and 1), and Color (ranging from 240 to 255). There is also a label field with 3 values (high, medium, low). Second, "Milk quality" dataset by Yrohit also consisted of 1059 rows and 7 fields similar to the previous dataset, with similar values, and includes a label field with 3 values (high, medium, low). Finally, "Milk Quality Prediction" dataset by Harini (2023): This dataset has a similar structure with 1059 rows and 7 fields, similar to the above datasets, along with a label field with 3 values (high, medium, low). All these datasets are related to predicting milk quality. However, only one dataset meets the requirements of the research, including numerical data, clear classification, and multiple fields to obtain objective results. Among them, the "Milk Quality Prediction" dataset by Rajendran was selected for the research as it was updated before the other two datasets and had been verified by the scientific research community for accuracy.

The Milk Quality Prediction dataset by Shrijiayan Rajendran was last updated on 02/08/2022. The normalized dataset consists of 8 features, totaling 1060 instances, including the following parameters in the given format: <Predicted Label> 1:<pH> 2:<Temperature> 3:<Taste> 4:<Odor> 5:<Fat> 6:<Turbidity> 7:<Colour>. These indices are derived entirely from the data source mentioned above and have been validated by experts in this field. Each feature will have the following indices:

- pH: Represents the pH level, should be entered for prediction within the range of 0 – 10, with the lowest value being 3 and the highest being 9.5.

- Temperature: Describes the temperature of the milk, should be entered for prediction within the range of 0 – 100, with the lowest value being 34 and the highest being 90.

- Taste: Indicates the taste of the milk, should be entered for prediction as 0 (poor) or 1 (good).

- Odor: Indicates the odor of the milk, should be entered for prediction as 0 (poor) or 1 (good).

- Fat: Represents the fat content of the milk, should be entered for prediction as 0 (low) or 1 (high).

- Turbidity: Indicates the turbidity of the milk, should be entered for prediction as 0 (low) or 1 (high).

- Colour: Describes the color of the milk, should be entered for prediction within the range of 240 – 255, with the lowest value being 240 and the highest being 255.

The downloaded dataset can be referred to as raw data, which cannot be directly used for this research due to the need for extensive transformation from various parameters to expressed characteristics. Raw data must undergo preprocessing before it can be applied to the training process of the software. From this raw data, unnecessary information such as serial numbers and IDs will be removed. These details are not essential for the training process of the program. Next, the labels will be transformed: labels like "high," representing high milk quality, will be converted to the number "1"; labels like "medium," representing "average" quality, will be converted to the number "2"; and the remaining labels, denoting "low" quality, will be transformed into the number "3." The remaining attributes like pH (pH level), Temperature, Taste, Odor, Fat, Turbidity, and Colour will be kept unchanged as they are already in numerical form and do not require conversion. All of these conversion processes will be carried out entirely using Microsoft Excel software.

After transforming the raw data into standardized format for use in the software, the dataset will have 7 main fields: pH, Temperature, Taste, Odor, Fat, Turbidity, and Colour, along with 1 label field containing the converted values of 1 (high), 2 (medium), or 3 (low), and a total of 1,060 rows. The dataset must be saved in a file with the ".csv" extension. Each data point should be entered in a single cell, and the parameters as well as the label must be separated by a comma ",". As per the system's requirement, the data's parameters should be placed before the label, and the label should be placed at the end. The data will be split into training and testing sets with a ratio of 70% and 30% respectively. After splitting the data, it will be saved to the computer. Next, the dataset will be split into batches for running the algorithms. The batch size will be set to 70% * n or 70% * n/2 with a 70% train/test ratio, where n represents the total number of data points.

Currently, fundamental databases are facing a significant challenge: their ability to adapt to the evolving nature of data over time. This issue arises because classical algorithms, on which current databases are often trained, can only undergo training once and need to be retrained from scratch when data changes occur. For instance, when a database is trained to create a model, with the arrival of new data, the database must start anew to generate a new model capable of accurate predictions. However, in modern data environments, data changes continuously over time, necessitating the ability to adapt rapidly to these changes.

The essence of "Concept Drift" involves the evolution and adaptation of models in response to changing environments. In non-stationary settings, the sliding window approach becomes pivotal, ensuring that models stay updated while handling variations effectively. This is especially important in domains where changes can happen rapidly or unpredictably, requiring agile methodologies to adapt and maintain performance over time.

The term "Concept Drift" has been widely used in many real-problem with the food testing. Indeed, the concept of drift forms the basis for both gradual and continuous changes and the "forgetting" of previous situations. However, challenges in non-stationary environments arise from the fact that circumstances can change rapidly or slowly, be forgotten, and even instances where knowledge reappears after it has disappeared. In such complex cases, the situation of "dual dilemmas" regarding stability or adaptation holds true. It's important to note that these approaches are not universally "incremental approaches”. In the fields, three types of proposed methods are listed as solutions to this issue:

Sliding Window Approach: Considering the evolution of concepts in non-stationary environments can be accomplished by utilizing a "sliding window" method, such as the FLORA approach. This principle involves updating the model at each time point using the most recent training data, defined by a window of time or a number of data instances. This approach can perform re-classification within a "group" type (on data selected by the temporal window) or update models if online learning methods allow. In this case, the "forgetting" (as mentioned earlier) is automatically managed by this learning method. This method generally involves three steps: 1) detecting concept changes by using statistical tests on different windows; 2) if an observed change occurs, selecting representative and recent data to adjust models; 3) updating models. The window size is predetermined by the user. The key point of these methods is to determine the window size. Most methods employ a fixed-size window that is configured for each practical scenario.

Dataset Temporariness: The dataset is considered temporary due to the emergence of many new standards, resulting in increased diversity that may lead to future changes in the dataset's classification. Consequently, this dataset can be classified as belonging to an unstable environment. Therefore, traditional machine learning algorithms used in static environments cannot be applied because conventional methods do not allow data augmentation. Instead, advanced algorithms capable of handling data in dynamic environments must be employed.

To address this issue, the concept of continuous learning in non-stable environments has been proposed as an alternative solution. Continuous learning allows databases to learn from data continuously in changing data environments, facilitating the update and adjustment of their predictive models. This enables databases to adapt to data changes and enhance the accuracy of their predictive models. Based on these characteristics and drawing from the history of research on various methods, the upcoming problem will utilize the sliding window approach. The sliding window method is the most suitable in this case, and it will be combined with the Extra Tree Classifier algorithm.

This approach enables databases to continuously adapt and update their models as new data arrives, ensuring that the predictive models remain accurate and relevant even as the data environment changes. This is particularly important in scenarios where data shifts are frequent and unpredictable, making it crucial for databases to maintain their performance and adaptability over time.

3. RESULTS AND DISCUSSION

The Sliding Windows approach, is a method commonly used to address the issue of concept drift in non-stationary environments. The key idea is to adapt the model continuously to the changing nature of the data over time. This method is often applied using techniques like FLORA. The principle involves updating the model at each time step using the most recent training data within a defined time window or a certain number of data points. This approach can involve reclassifying the "group" (based on the temporarily selected data) or updating the model if online learning techniques are applicable. In this case, the process of "forgetting" (as mentioned above) is managed automatically through this learning method. The Sliding Windows approach typically includes three main steps:

- Detecting Concept Changes: This is done by employing statistical tests on different sliding windows to identify changes in the concept.

- Adapting to Changes: If a concept change is detected, representative and recent data points are selected to adjust the models accordingly.

- Updating Models: The models are then updated based on the chosen data points to account for the concept drift.

The size of the sliding window is determined by the user and is a key parameter of these methods. Most approaches use a fixed-size window that is configured according to the specific characteristics of the real-world problem. The Sliding Windows approach is effective in capturing changes in the underlying data distribution and allows the model to adapt to these changes over time. This is particularly valuable in scenarios where the data is not stationary and concepts evolve. The ability to dynamically adjust the model's parameters and structure based on the most recent data helps maintain the model's relevance and predictive accuracy in a changing environment.

All achieved results will be based on Balanced Accuracy, which is a measure that can be used when evaluating the performance of a binary classifier. It is particularly useful when classes are imbalanced, meaning one class appears much more frequently than the other. Using Balanced Accuracy is much more complex than traditional accuracy. Regular Accuracy simply calculates the percentage of correctly classified data points based on the total available data. While this initially works well, when data is heavily skewed, the accuracy's reliability diminishes. To address this, a calculation method was devised based on true negative and true positive rates, which can compute true negative rate, true positive rate, false negative rate, and false positive rate. After obtaining these values, the Balanced Accuracy formula can be used to compute the most accurate and optimized percentage. The formula for Balanced Accuracy used in these experiments is as follows:

- First need to consult the matrix table: Absolutely, referencing a confusion matrix is a crucial step when evaluating classification results. A confusion matrix provides a clear overview of the performance of a classification algorithm by displaying the true positive, true negative, false positive, and false negative counts for each class. It's a helpful tool to understand how well your model is performing, especially in situations where classes might be imbalanced.

- Must have:

TPR (true positive rate) = TP/(TP+FN)

TNR (true negative rate) = TN/(TN+FP)

- After obtaining the above two parameters, Balanced Accuracy is calculated using the formula:

Balance Accuracy = (TPR+TNR)/2

Based on the data presented in the chart, we can observe the following average ratios and the final step of the Extra Tree Classifier algorithm:

The performance of the Extra Tree Classifier appears to be quite stable, with an average ratio across all steps reaching 60%, specifically 70.12%. The average accuracy rate of the final step also reaches 75%. This indicates that the Extra Tree Classifier is capable of making reasonably accurate predictions in various scenarios, with relatively good accuracy. Therefore, the Extra Tree Classifier algorithm seems suitable for addressing the milk quality prediction problem, as it demonstrates the ability to provide fairly accurate predictions with a relatively high level of accuracy (Figure 2).

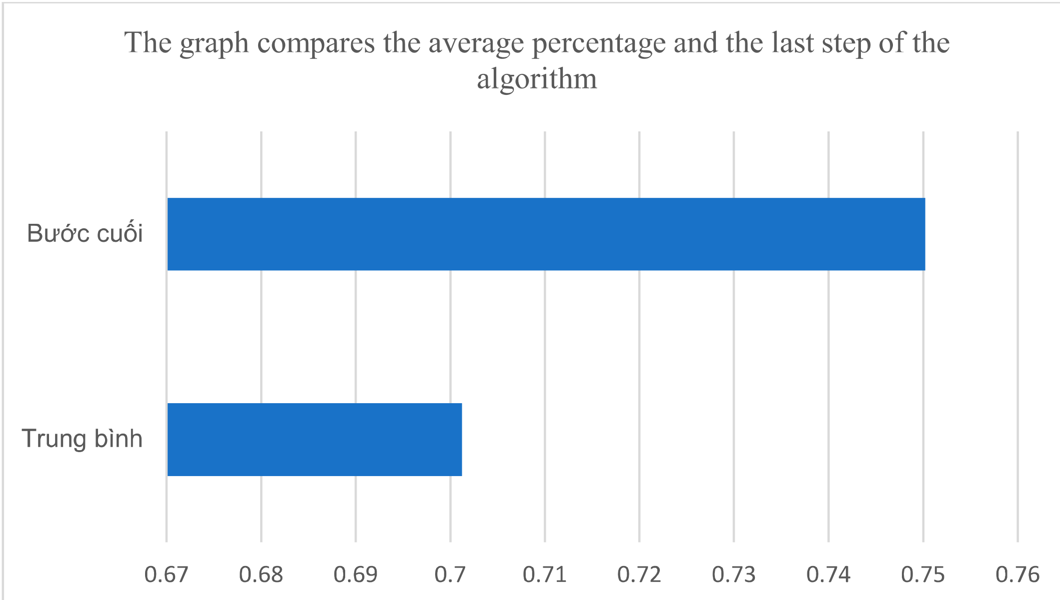


Figure 1. The graph compares the average percentage and the last step of the algorithm (Extra Tree Classifier)

Sure, you can compare the performance of the Extra Tree Classifier algorithm in more detail by analyzing the accuracy progression for each step. The accuracy progression chart of the Extra Tree Classifier algorithm provides a clearer and more comprehensive view of its performance. This chart can help you evaluate the algorithm's behavior and effectiveness throughout its execution.

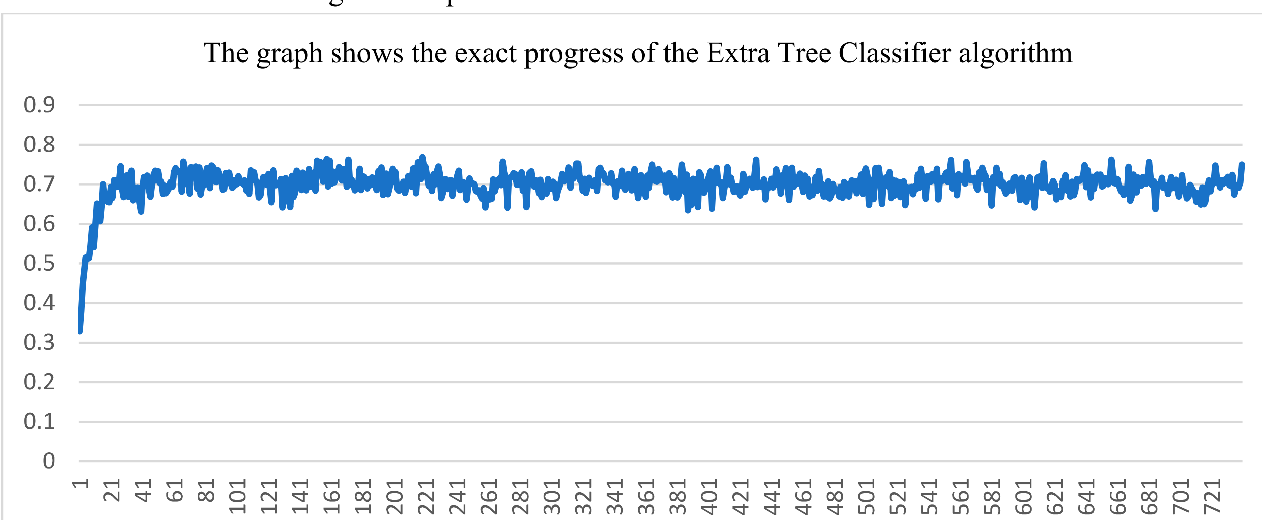


Figure 2. The graph shows the exact progress of the Extra Tree Classifier algorithm

Based on your description of the accuracy progression chart for the Extra Tree Classifier algorithm, it seems that the algorithm starts with a relatively low accuracy of 37% but gradually improves its stability. The most noticeable improvement occurs from step 1 to step 20. In the beginning, the algorithm's average accuracy is only 31%, but by step 20, it increases to 70%, and then reaches a stable level. The accuracy of the Extra Tree Classifier seems to show relatively stable fluctuations ranging from 60% to a peak of 76.22%. The highest accuracy is achieved at step 219, with a rate of 76.86%. This is a substantial value that can be compared favorably to many other algorithms. Other steps with high average accuracy include step 172 (76.21%), step 270 (75.67%), step 318 (75.3%), step 431 (76.18%), and step 555 (76.12%). Furthermore, you mention that the most significant growth in accuracy occurs between steps 421 and 431. This information provides valuable insights into how the algorithm performs and where its strengths lie.

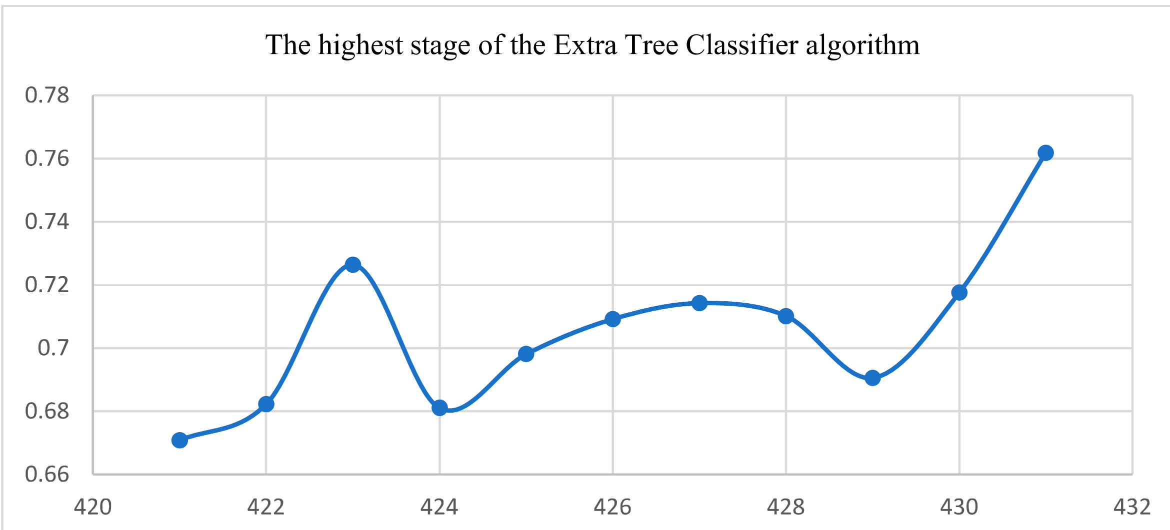


Figure 3. The highest stage of the Extra Tree Classifier algorithm (Batch 421-431)

Observing the chart, it is evident that there is stability and gradual improvement in the accuracy progression of the Extra Tree Classifier algorithm. The accuracy consistently reaches above 60% by the step with the highest rate. This level of stability is quite satisfactory compared to other algorithms. Despite occasional periods of less favorable results during data processing, the Extra Tree Classifier consistently stabilizes relatively quickly, as depicted in the chart below:

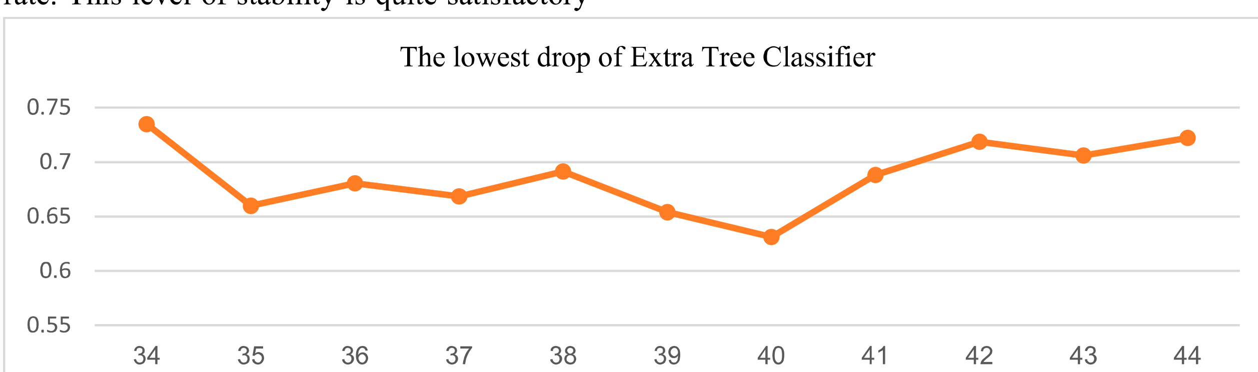


Figure 4. The lowest drop of Extra Tree Classifier (Batch 34 -44)

At step 40, the Extra Tree Classifier achieves its lowest accuracy rate of 63.11%. However, by step 44, the algorithm quickly regains stability with an accuracy of 72.2%, and it gradually stabilizes in the subsequent steps. Although this period represents the lowest accuracy phase for the Extra Tree Classifier, the accuracy remains above 60%, indicating a reasonably stable phase. The discrepancies in accuracy values are not too large during this phase. For example, in the range from step 35 (65.96%) to step 39 (65.4%), there is a slight drop in accuracy from step 34 (73.47%) to step 35 (65.96%), followed by stability from step 35 to step 38 (69.13%), a deep decrease to 63.11% at step 40, and then an increase back to 72.2% at step 44.

From the comparative charts, it can be observed that the Extra Tree Classifier algorithm demonstrates relatively good stability with limited fluctuations in the average accuracy rates across steps. The accuracy of the algorithm is also notably high, ranging from 60% to 80%. However, it's important to note that each algorithm has its own strengths and weaknesses. Choosing an appropriate algorithm depends on the specific requirements of the problem and the characteristics of the dataset at hand.

Application installation:

Based on the final results presented earlier, the Extra Tree Classifier algorithm will be chosen as the method to address the issue of milk quality diagnosis in the application. The system has been completed and meets all the initial requirements, and it is divided into distinct sections to serve both regular users and developers. Regular users will have an interface to perform milk quality diagnosis. On the other hand, developers will have more interface pages to work with and develop within the system. These interface pages include the login page, data training page, diagnostic function testing page, model list display page, and settings page. The system has a high practical applicability, especially in the context of increasing milk consumption in Vietnam. Over time, regulations regarding milk quality may change due to various objective reasons. Therefore, the dataset will also undergo significant changes. With the Extra Tree Classifier algorithm, the system is capable of effectively addressing and adapting to these changes.

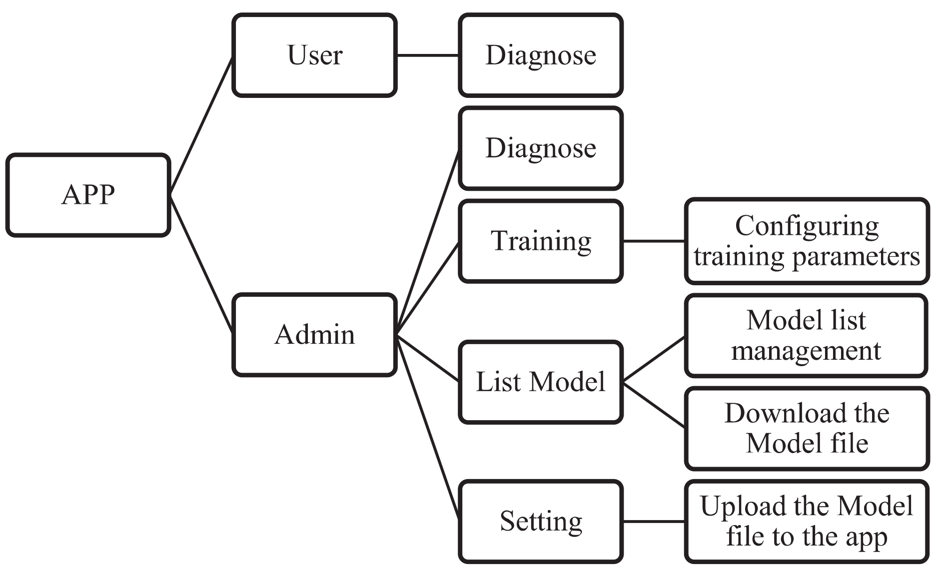


Figure 5. Use case diagram of the system

The section for predictors:

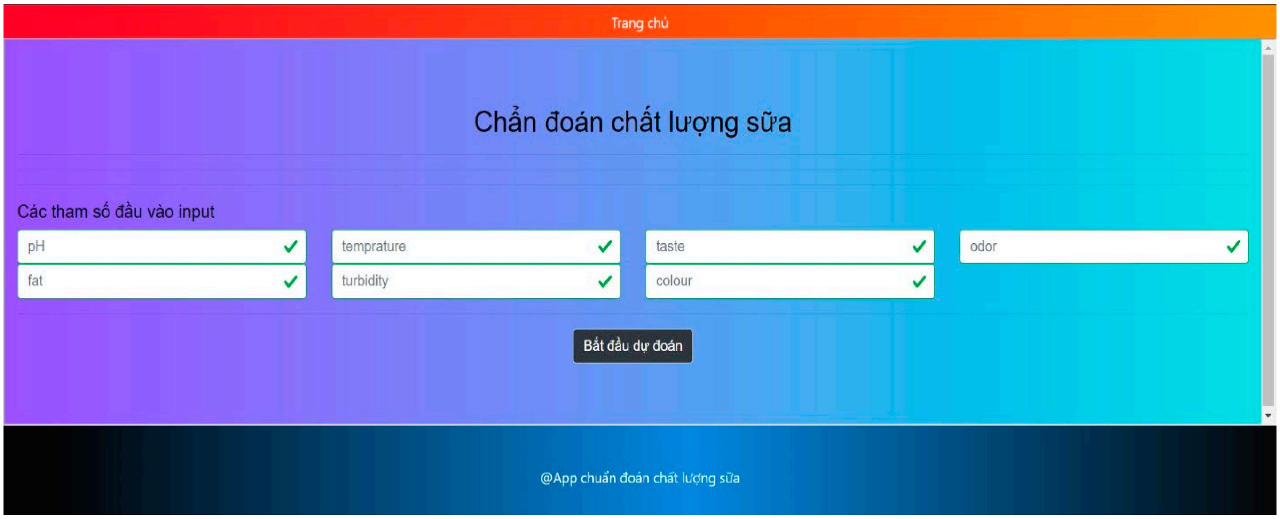


Figure 6. Main interface of Milk Quality Diagnostic system

The main interface of the app is where users can perform the diagnosis. Users need to input complete information into various input fields. The required information includes the 7 attributes of milk quality: pH level, temperature, taste, turbidity, fat content, odor, and color. After providing these inputs, users can press the "Start Prediction" button to receive the prediction results.

In this project, the system has been packaged as a .ZIP file named "ChanDoanSua-main.zip". After users download and extract it, they will find a folder named "BaoCaoThucTap". Inside this folder, there are files to run the program. Below are the installation instructions:

The system requires a computer that is always connected to the internet and meets the minimum configuration: Windows 10, 2GB RAM, and 10GB hard drive space. To run the software, follow these steps:

- Go to the "SETUP" folder and install Python. The version used is Python 3.9.9 (python-3.9.9-amd64.exe).
- Install the required libraries by running the file "CaiThuVien.bat".
- Run the "RunServer.bat" file to start the program.
- Open a web browser and navigate to the address 'http://127.0.0.1:8000/'.
- The program will run on the default port: "http://127.0.0.1:8000/".

The program can also be deployed on a web platform. This is the default installation process.

4. CONCLUSION

The project has achieved several significant outcomes. Firstly, it involved a comprehensive study of milk quality trends both globally and in Vietnam, providing insights into the crucial role of quality assessment. The project successfully managed the database, selecting and transforming it to meet the project's specific needs. Additionally, the project delved into the workings and practicality of the Extra Tree Classifier algorithm, effectively utilizing it to address the problem at hand. An essential aspect of the project was the meticulous comparison of algorithm stability, resulting in the identification of the most suitable approach. This culminated in the development of a diagnostic application featuring an intuitive and user-friendly interface. Looking ahead, the project envisions further enhancements. This includes refining the user interface for better aesthetics and usability, expanding the application's capability to diagnose diverse datasets from various domains beyond milk quality, and refining the algorithm for heightened accuracy. The project's context is rooted in the Vietnamese milk market, acknowledging concerns about subpar milk quality affecting both export potential and public health. By leveraging the Extra Tree Classifier algorithm alongside sliding window techniques, the system adapts to dynamic data changes. The software, developed as a web-based application, ensures an accessible interface that caters to developers and end-users alike. The visual system model encapsulates user interaction and developer installation procedures, using commonly employed software tools.

REFERENCES

[1] Elwell (2006*). Use of microfiltration to improve fluid milk quality.* https://www.sciencedirect.com/science/article/pii/S002203020672361X

[2] Marth (1981). *Assuring the quality of milk.* https://www.sciencedirect.com/science/article/pii/S0022030281826793

[3] Marth (1972). *The thirteenth edition of standard methods for the examination of dairy products.* https://www.sciencedirect.com/science/article/pii/S0362028X23022202

[4] Tường Vi (2019). *7 nguyên tắc vàng mang đến từng hộp sữa tươi an toàn và chất lượng.* https://vov.vn/suc-khoe/7-nguyen-tac-vang-mang-den-tung-hop-sua-tuoi-an-toan-va-chat-luong-964919.vov

[5] VIRAC. (2017). *Tổng quan ngành công nghiệp sữa.* https://investvietnam.gov.vn/vi/nghanh.nghd/15/sua-va-cac-san-pham-sua.html

[6] Viện Kiểm Nghiệm An Toàn Vệ Sinh Thực Phẩm Quốc Gia. (2022). *Kiểm nghiệm sữa và sản phẩm sữa.* https://nifc.gov.vn/ky-thuat-chuyen-mon/kiem-nghiem-sua-va-san-pham-sua-post1484.html

[7] Padmaja, B. (2021). *Prognosis of Vitamin D deficiency severity using SMOTE optimized Machine Learning Models*. https://turcomat.org/index.php/turkbilmat/article/view/8442

[8] Bachmann, H.P. (2023). *Classification of cheese varieties from Switzerland using machine learning methods: Free volatile carboxylic acids.* https://www.sciencedirect.com/science/article/pii/S0023643823006746.

[9] Mujumdara, A. (2019). *Diabetes Prediction using Machine Learning Algorithms.* https://www.sciencedirect.com/science/article/pii/S1877050920300557