MILK QUALITY DETECTION USING MACHINE LEARNING ALGORITHMS

Abhishek Singh Yadav , Belo Abhigyan , Rahul Kumar Manjhi

Problem Statement

How can we develop a machine learning model to effectively monitor milk as a perishable commodity, aiming to minimize financial losses and mitigate health risks by accurately predicting and preventing spoilage?

Abstract:

In the digital age, machine learning (ML) algorithms are critical for data analysis and decision-making in a variety of industries. The food industry, as one of the most important sectors, stands to benefit greatly from rapid and accurate assessment of product quality. Milk being a perishable commodity that must be closely monitored to avoid financial losses and reduce the health risks associated with spoilage. This study investigates the assessment of milk quality using machine learning techniques, leveraging a dataset sourced from the Kaggle repository. Employing seven distinct features, including pH, temperature, taste, odor, fat content, turbidity, and color, milk quality classification was performed. Among numerous machine learning algorithms applied in our study, the two widely-used algorithms, Neural Network (NN) and Adaptive Boosting (AdaBoost), were proven to give exceptional results for classification estimation. The results from above were visualized and compared, showing that AdaBoost performed better than NN, outperforming it with a 99.9% classification accuracy, outperforming NN's 95.4%. Furthermore, the milk samples were categorized into low, medium, and high grades, and an Artificial Neural Network (ANN) model exhibited outstanding accuracy in classification, outperforming all other previous methods except AdaBoost. This study demonstrates how machine learning (ML) algorithms can improve quality control procedures in the food industry and offer opportunities for additional refinement through the use of deep learning techniques.

Introduction:

The Food Safety and Standards Authority of India (FSSAI) conducted a National Survey on Milk Adulteration in 2011, which revealed that water was the most common adulterant in Indian milk, followed by detergent. To address this issue and safeguard the health of consumers, the Indian government introduced the Prevention of Food Adulteration Act (PFA Act) in 1954, which came into effect on June 1, 1955 [1]. The production, sale, and distribution of contaminated and poisonous foods are prohibited by law. Despite these efforts, fortification persists due to a shortage of well-trained personnel and inadequate laboratories. To improve the quality of food products, various studies have been conducted using data from reputable sources and employing machine learning techniques. This shows that it is possible to use the neural network technology to ensure that milk produced is of high quality based on numerical data. The potential of new technologies will be realized if new approaches are identified to manage milk quality using neural network technologies [2], thereby saving resources and preventing health hazards from occurring. Using numeric data, machine learning methods such as neural networks and decision trees can be utilized to develop models that embody the properties of both high-grade and adulterated milk samples. Through precise detection, the expenses associated with treatment can be diminished, the spread of illness caused by substandard milk can be curtailed, and the quality of milk can be sustained. Efficient detection not only reduces costs but also ensures the well-being of cattle by detecting diseases or infections in advance. By using a computational model, farmers can apply appropriate treatment plans quickly, preventing the disease from infecting healthy cattle [3]. The actual motive of this research is to improve the precision of milk predictive analysis. To achieve this, machine learning (ML) techniques along with deep learning methods are employed instead of traditional methods such as cluster analysis and discriminant analysis. The data is divided into training and testing sets, which are used to train the ML algorithms. This enables the computer to refine its predictions by learning from previous results.

In our study, we thoroughly evaluate various machine learning algorithms for classifying milk quality. These algorithms include regression techniques and classifiers like Support Vector Machines (SVM),AdaBoost, K-Nearest Neighbors (KNN), and Random Forest (RF). Additionally, we implement Artificial Neural Networks (ANN) and compare their performance with the other algorithms used previously. On comparing the results we found out that the ANN model achieves an excellent balance of computational efficiency, accuracy, and stability, making it the most suitable choice for milk quality classification.

Background :

Accurately classifying milk samples is essential for ensuring the quality of dairy products supplied from farms. The presence of various chemical compounds and physical factors that affect milk quality makes it challenging to establish a reliable and consistent classification approach. Extensive research has been carried out to find an effective solution to this milk classification problem.

Xiao et al. (2019) set up a random forest model, LR model, and AdaBoosting model and performed tests to find the most appropriate classification model. In the study, they showed that the color, aroma, and taste of milk as attributes can easily recognize the quality of milk. According to the results obtained, the success rate was found to be 96.8%.

In a recent study by Olcay Polat (2021) [4], the use of an information fusion framework was explored as a means of classifying raw milk samples. The aim was to develop a suitable approach to categorize raw milk into multiple classes based on important criteria. The study results showed that pH, sH and somatic cell count were significant factors in determining the quality of raw milk. In the field of dairy processing, Multi-criteria Decision-Making (MCDM) and Analytic Hierarchy Process (AHP) techniques were implemented, but their computation complexity was high. To address this issue, Pegah Sadeghi Vasafi (2021) [5] utilized KNN and SVM classification techniques to identify anomalies in the milk processing process. The fat content and temperature were considered as features and the accuracy achieved was 81.4% for SVM and 84.8% for KNN. Despite the success of these two techniques, a higher accuracy could have been achieved by incorporating additional classifiers. W Habsari (2021) [6] aimed to develop a smart grading system for milk quality classification in the dairy industry using ANN and K-means models. The factors affecting milk quality, such as pH, temperature, odor, turbidity, color, fat and taste values, were considered in the study. The accuracy obtained was 98.74%, with color and temperature being the attributes used for grouping. In 2013, Shailesh Chaturvedi [7] employed ANN model to predict the milk amount and constructed two hidden layers of the network. The model's performance was compared to actual experimental data and showed that ANN could be used to determine the future performance of dairy cattle based on early traits expressed. The network consisted of four input layers, two hidden layers, and one output layer. 60% of the data was used for training and 20% each for testing and verification. However, the error could have been reduced and the correlation coefficient maximized closer to 1.

Earlier, in 2002, A.P. Kominakis [8] studied the use of ANN for predicting the yield of milk and found that network specialization improved the prediction ability, resulting in a high correlation between observed and predicted values (0.87 to 0.97). However, the standard deviation of the data was misjudged, leading to an unexpectedly high standard deviation (s=91.55).

In this study, AdaBoost and Neural Network algorithms were applied on the Milk Quality dataset shared in Kaggle data storage and the success of estimating milk quality was compared.According to the results obtained, it was seen that the AdaBoost algorithm made a highly successful quality classification.

Materials and Methods

Data Collection

The data set for the proposed system Collected from Kaggle repository. This dataset consists of 7 independent features as shown in the table1. These parameters are used to predict analysis of the milk. Grade (Target) of the milk which is categorical data where Low (Bad) or Medium (Moderate) and High are three different classes.The total number of records present in the dataset is 1059 rows, and 8 columns Out of all features, 7 are categorical and 1 is numeric.

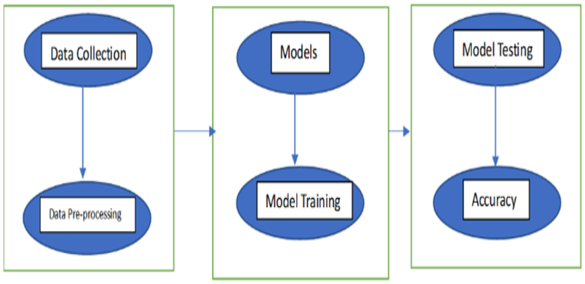
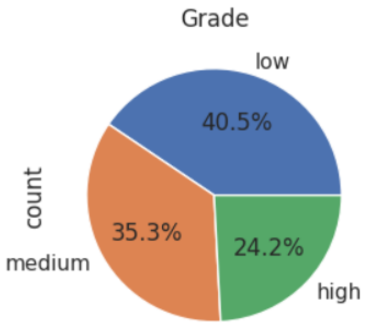


Table 1. Categorical and Numerical data

Categorical Data	Numerical Data
Grade	pH
	Odor
	Temperature
	Taste
	Fat
	Color
	Turbidity

Dataset Description

The dataset utilized in this study to predict milk quality and grade was sourced from Kaggle, a reputable open-source data repository. It consists of 1059 rows and 9 columns, as outlined in Table 2. The data, manually collected, provide crucial insights for developing machine learning models aimed at predicting milk quality. The dataset comprises seven independent variables: pH, Temperature, Taste, Odour, Fat, Turbidity, and Colour, all of which are instrumental in determining milk quality or grade.

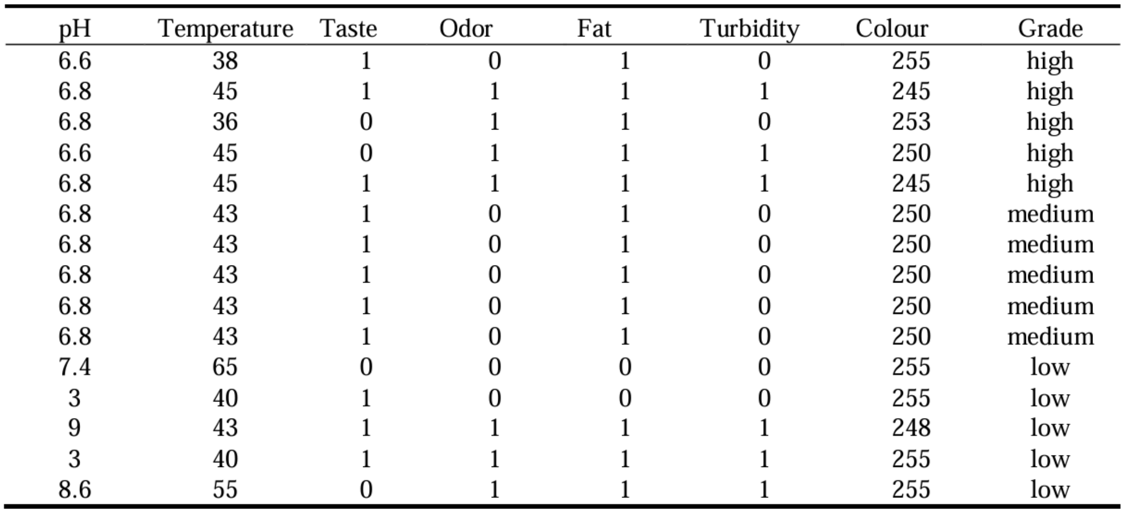


Specifically, the pH column records the potential of hydrogen values of the milk samples, ranging from 3 to 9.5. This parameter is vital for detecting impurities and signs of infection in the milk. The Temperature column denotes the temperature at which the milk is maintained, ranging from 34 to 90 degrees Celsius, crucial for preserving milk quality by preventing bacterial growth. Taste, Odour, Fat, and Turbidity columns encode binary values (1 or 0) representing good or bad quality indicators for these attributes.

Furthermore, the Colour column serves as an indicator of physico-chemical changes in the milk, with values ranging from 240 to 255, reflecting variations in color due to factors such as suspended fat globules and other constituents. Lastly, the Grade column categorizes milk quality into three classes: high, medium, and low, based on the observed attributes. Notably, no data pre-processing or cleansing was undertaken due to the absence of null values in the dataset.

Table 1 shows the values of 15 randomly selected milk samples (Kaggle,2022)

Table 1. Milk Quality dataset and classification (Kaggle, 2022)



pH	Temperature	Taste	Odor	Fat	Turbidity	Colour	Grade
6.6	38	1	0	1	0	255	high
6.8	45	1	1	1	1	245	high
6.8	36	0	1	1	0	253	high
6.6	45	0	1	1	1	250	high
6.8	45	1	1	1	1	245	high
6.8	43	1	0	1	0	250	medium
6.8	43	1	0	1	0	250	medium
6.8	43	1	0	1	0	250	medium
6.8	43	1	0	1	0	250	medium
6.8	43	1	0	1	0	250	medium
7.4	65	0	0	0	0	255	low
3	40	1	0	0	0	255	low
9	43	1	1	1	1	248	low
3	40	1	1	1	1	255	low
8.6	55	0	1	1	1	255	low

Data Pre-processing

The first step of preprocessing involved calculating the missing value in the data. It is found that there is not a single missing value for any of the features. The label is encoded in the following step. Since a computer cannot comprehend the value of the attribute in the problem, label encoding is used to convert the values in this case to category integer values. This dataset's "Grade" feature employs label encoding. The final step is to scale the feature values. This is where the Min-Max scaling or z-score scaling is used. Equation 1 illustrates the use of max and min values for scale in min-max scaling. Before model fitting, firstly feature-wise normalization, such as Min-Max. Scaling is typically employed to improves model convergence, and prevents certain features from overshadowing others based solely on their magnitude.Temperature is right-skewed with 2.216739 and color is left-skewed with -1.024902. We can use PowerTransformer to get rid of skewness. PowerTransformer also has standardize argument. We can set it to True to accomplish our second task: scaling.

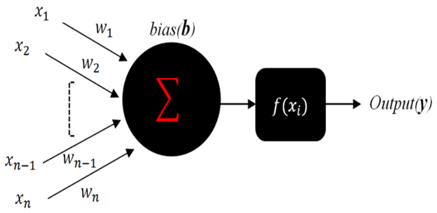
$$m = (x - x_{min}) / (x_{max} - x_{min}) \qquad (1)$$

where
x= feature value
x_{min}= minimum value of feature
x_{max}= maximum value of feature

Fig1. System Architecture

Methods Applied:

I. Artificial neural network-based milk quality detection



Neural networks are generally examined in 5 basic structures. These are Inputs, weight coefficients, bias (constant), activation function and output value (Balaban and Kartal, 2018). Figure 2 shows the model of the artificial neural network.

The mathematical expression of the neural network is shown in equation 1. is the output value, is the weight coefficient, is the input value, and is the constant coefficient.

$$y = w_i x_i + b$$

An input unit is passed with some weights attached to it to the hidden layer in the first step. The number of hidden layers is not limited. X1, X2, X3 ... Xn are the inputs. Neurons are present in each hidden layer. Neurons are connected to all inputs. As soon as the inputs are passed on to the hidden layer, all computation occurs there. The Calculations in hidden layers are performed in two steps as follows: To begin with, each input is multiplied by its weight. Variables are weighted according to their gradients or Coefficients. This indicates how strong a particular input is. Adding the bias variable comes after assigning weights. Adding bias to a model will ensure that it fits as accurately as possible as shown in Equation above.

A linear equation denoted by y here is then activated by applying an activation function. Before sending input to the next layer of neurons, the activation function transforms it nonlinearly. Activation functions instill nonlinearity in models. There are several activation functions such as Sigmoid, ReLu, LeakyReLU, Tanh, and SoftMax which can be applied on our model.

Every hidden layer undergoes the above process. After passing through each hidden layer, which is our output layer, we arrive at the final output. We call this process forwarding propagation. The error is calculated based on the predicted output and the actual output, which is the difference between the two. We use Back Propagation when the error is large in order to minimize it.

II. Gradient boosting based milk quality detection

Gradient boosting decision tree is an integrated boosting algorithm based on CART learner. The purpose of its algorithm in each round of iteration is to minimize the loss function of the current learner so that the loss function always decreases along its gradient direction, and the final residuals approach 0 through continuous iteration, adding up all the tree results to get the final prediction.

III. XGBoost based milk quality detection

Extreme gradient boosting algorithm is an improved version based on GBDT, which is not sensitive to input requirements and is widely used in the industry. Compared with the general GBDT algorithm, XGBoost uses the second derivative of the loss function about the function to be sought, adds a regularisation term to prevent overfitting, and samples the attributes when constructing each tree. It has fast training speed and high accuracy and fitting eect, etc.

IV. Support vector machine-based milk quality identifier

SVM is a supervised machine learning algorithm mainly used for classification and regression tasks. The core idea is to find the hyperplane that best classifies the data, or in the case of regression, fits the data most efficiently. SVM does this by maximizing the distance between the hyperplane and the closest data points in the two classes, called support vectors. SVM can effectively handle high-dimensional data and is therefore suitable for predicting milk quality based on multiple parameters. SVM can handle nonlinear relationships between parameters using kernel functions such as radial basis functions (RBF), polynomials, and sigmoid. SVMs are less prone to overfitting, especially when margins are chosen carefully. This means that data that is not yet visible can be better summarized, which is crucial for accurate predictions of new milk samples.

V. Quality assessment of milk using Random Forest

Random Forest is an ensemble learning technique that builds multiple decision trees during training and outputs a mean prediction (for regression tasks) or class (for classification tasks) for each tree based on unknown data. The decision trees that make up the "forest" it creates are usually trained through the use of "bagging" techniques. Numerous features in data sets can be managed by random forests, which can also assess how significant each feature is in predicting the quality of milk. In practice, it's possible that some milk samples lack certain measurements. Random forests are capable of producing precise predictions even in the presence of missing values. The tendency of decision trees to overfit is one of their drawbacks. Nevertheless, by employing multiple trees and averaging their output, random forests can achieve better generalization.

In the case of the classification problem, the equation for Gini will be as demonstrated in Equation below :

$$X = 1 - \sum_{i=1}^{n} t_i$$

Here, X is the Gini to calculate with n representing the classes present and the relative frequency of the present class from the data set being observed. In place of Gini, we can also use the entropy method where we take X as the entropy to be calculated which helps to determine the way the tree branches from the decision tree. The equation gets modified as represented in the Equation given below;

$$X = \sum_{i=1}^{m} - l_i * \log_2 t_i$$

VI. K-nearest neighbor-based milk quality identifier

The k-nearest neighbors (KNN) is another efficient machine learning algorithm used for data classification. It helps in estimating the probability of whether a random data point will belong to one or the other group on the basis of checking the belongingness of the data points nearest to it. It can be used for solving both classification and regression problems but is more helpful in the case of a prior one [13]. While choosing the K value, we should try to pick up an odd value for it. There are several distance methods available with the help of which nearest or neighborhood points can be calculated. Using Euclidean Distance, the equation formed will be as stated in Equation stated below:

$$d = \sqrt{\sum_{i=1}^{n}(y1_i - y2_i)^2}$$

Where d is the Euclidean distance and (y1-y2) gives the shortest distance between two points. The method is used when data of high dimension is not to be considered. However in the case of high dimensional data and especially when it is a grid type of data, we use Manhattan Distance which is given as stated in the The equation below

$$d = \sum_{i=1}^{n} |y1_i - y2_i|$$

VII. AdaBoost machine learning algorithm based Quality Detection

By combining many rules, the correct prediction rule creation process is carried out with a machine learning approach. The AdaBoost algorithm was first applied in practice by Freund and Schapire in their study in 1996. They proved that this method is a new machine learning algorithm (Freund and Schapire, 1999). It is widely used in many fields with its strong classification ability (Wang et al., 2019). Adaboost algorithm can be applied to many classifier learning algorithms and strong classification results are obtained. In the AdaBoost algorithm, the “weight” value is calculated by analyzing all the data in the data set (Sun et al., 2006). With the AdaBoost algorithm, the input value x is divided into two classes as shown in the equation below. The y_ class values are shown as -1 and 1.

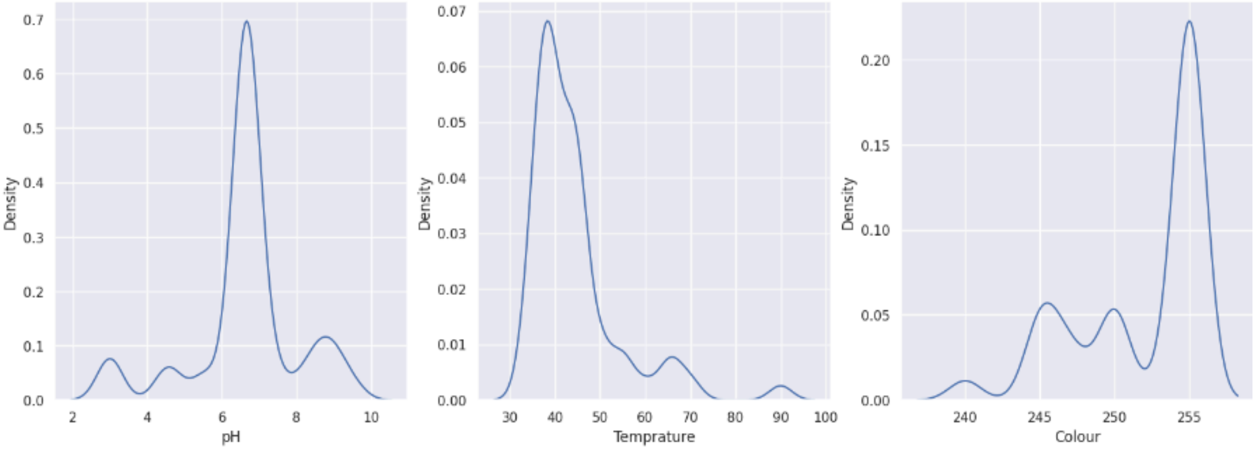
$$x_i, x_{i+1}, x_{i+2} \text{ and } y_n \in \{-1,1\}$$

Calculation of the weighting coefficient of n pieces of x data is shown in the equation given below:

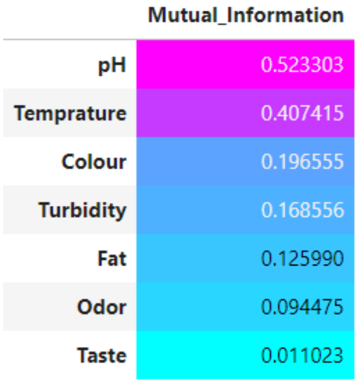
$$w_i = \frac{1}{n}$$

Observations

- There is a space in column named "Fat ". So we trimmed the extra whitespaces present in the column name.
- All features dtypes are integers except "pH" values are float, "Grade" values are object.
- Milk with higher and lower pH values seem to have low quality. All pH values are positive in range(3.0: 9.5).Majority of values are 6.7.
- All temperature values are positive and in range(34.0: 90.0) Majority of degrees - 45.0.
- All color values are in range(240.0: 255.0). Majority - 255.0.
- There are samples with bad taste but classified as high quality.

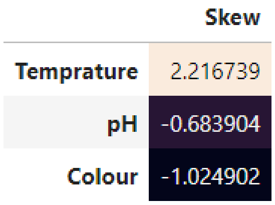


- Milk quality is medium to high if pH in range (6.59: 6.67)
- If temperature increases milk Quality decreases.
- pH value must be kept in a certain interval at least for medium quality.
- Higher temperatures could result in low quality.
- Color is not a very good feature to distinguish high and low quality milk.
- Good taste average is higher at high-quality samples.
- Fat is necessary for high-quality milk but it is not sufficient.
- Odor average is higher at high-quality samples (People like smelly milk)
- Turbidity could lead low-quality risk
- pH and Temperature will make a really good job at predictions. The others are also have significant mutual information scores.



	Mutual_Information
pH	0.523303
Temprature	0.407415
Colour	0.196555
Turbidity	0.168556
Fat	0.125990
Odor	0.094475
Taste	0.011023

- As data is not in a fixed range so feature scaling is needed for the contours will look more like circles and gradient descent can find a much more direct path.
- Temperature is right-skewed color is left-skewed.



	Skew
Temprature	2.216739
pH	-0.683904
Colour	-1.024902

Results and Discussion

In our study, it was seen that the pH value and temperature value of the milk play an important factor in determining the milk quality. It has been observed that the pH value of high- and medium-quality milk is between 6-7, and the temperature values are at most 45 degrees. In low-quality milk, it was observed that these two conditions did not occur at the same time.

As we implemented all the machine learning algorithms mentioned above, The entire dimension of the data set, 1059 rows by 7 columns, was taken for training, testing, and evaluation. We performed a data visualization exercise on the target variable, 'Grade', where each feature was thoroughly explored and analyzed.

Then we applied different models like SVM, RF,Adaboost,GBM ,XGB, DT and KNN. The different metrics evaluation sections like precision, F1-score, recall are mentioned in Table 3. In terms of classifier accuracy scores, Random Forest (RF), K-Nearest Neighbor (KNN), XGBoost(XGB), and Adaboost achieved a score of 0.9858, while Gradient Boosting(GBM), Decision Tree (DT), ANN and Support Vector Machine (SVM) showed scores of 0.9811, 0.9717, 0.967 and 0.9151 respectively. The precision score, which measures the number of positive class predictions that are actually positive, was used to determine the accuracy of the minority class in an imbalanced classification. The precision score increased with an increase in the minority class. The recall, which refers to the number of positive class predictions achieved, and the F1-score, which strikes a balance between precision and recall, also increased in the order of Random Forest (RF), K-Nearest Neighbor (KNN), XGBoost(XGB), Adaboost ,Gradient Boosting(GBM), Decision Tree (DT), ANN and Support Vector Machine (SVM). The F1 score, which combines both precision and recall, is a useful metric to achieve successful classification. The F1 scores obtained ranged from 0.9 to 0.99, indicating a successful classification of almost all observations into the correct class by the models.

Secondly, to build a neural network model, we proceeded with encoding the categorical data. Since our ‘Grade’ column had string values, the respective three classes, which were high, low, and medium, were assigned the numerical values 0, 1, and 2, respectively. Then the splitting of the data set was performed with a test value of 0.2.

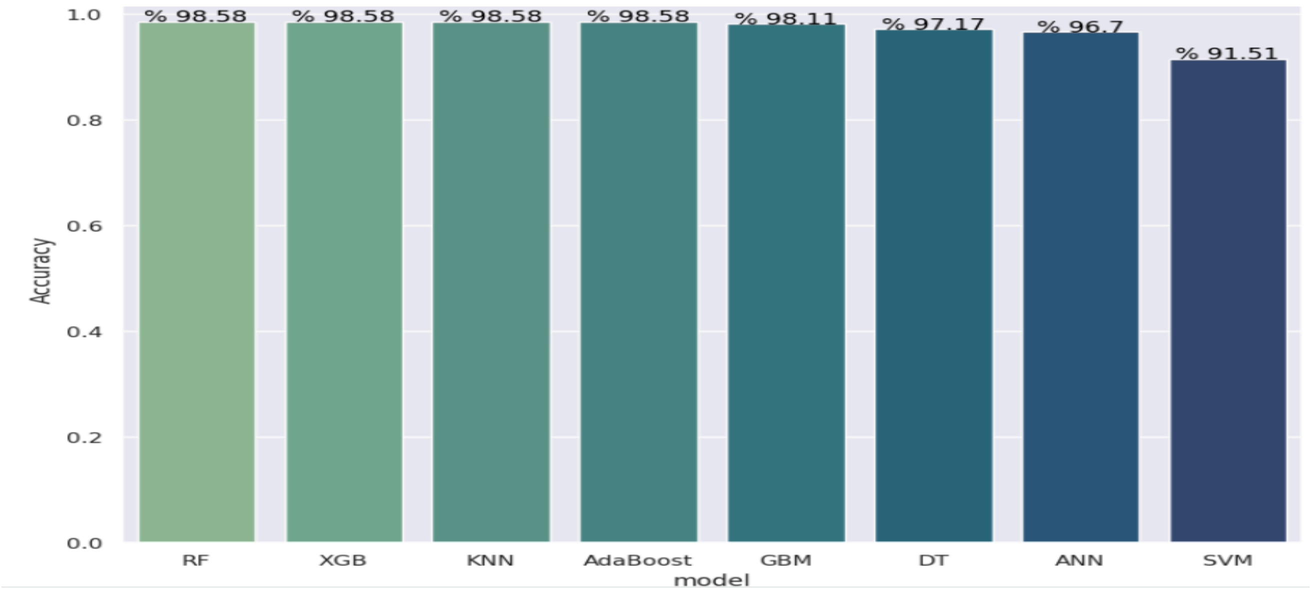
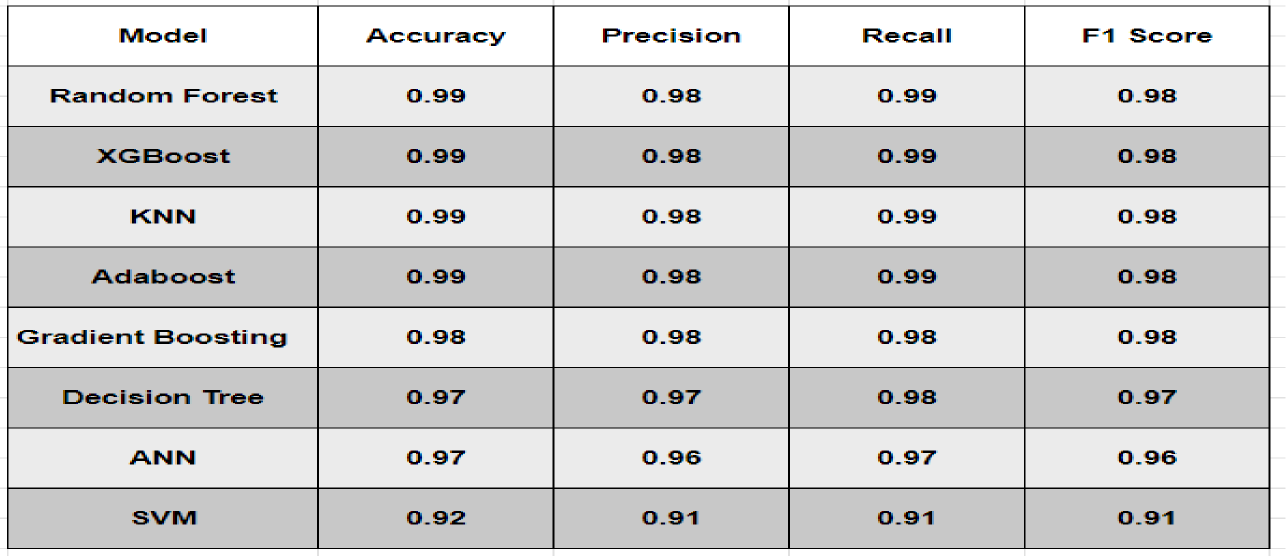


Table 3. Metrics Evaluation of models used



Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.99	0.98	0.99	0.98
XGBoost	0.99	0.98	0.99	0.98
KNN	0.99	0.98	0.99	0.98
Adaboost	0.99	0.98	0.99	0.98
Gradient Boosting	0.98	0.98	0.98	0.98
Decision Tree	0.97	0.97	0.98	0.97
ANN	0.97	0.96	0.97	0.96
SVM	0.92	0.91	0.91	0.91

Later, we proceeded with feature scaling, where the standard scalar was used, and finally, the ANN model was built. We had {20, 15, 10, and 3} layers, and the epoch value was set to 50. The first 3 layers had an activation function called ReLu, and the last output layer had SoftMax activation. For the artificial neural network, the batch size was set to 10, and adaptive moment estimation (Adam) was used as an optimizer. Furthermore, after setting the epoch value to 50, the artificial neural network model achieved an accuracy score of 0.967, demonstrating the reliability and efficiency of neural network models in dealing with classification problems compared to regression methods.

Conclusion

The paper employs a machine learning algorithm to predict milk quality based on influential factors, crucial for ensuring high-quality dairy products and preventing health risks. With increasing milk consumption and global production, efficient quality control becomes challenging. Machine learning offers a time-saving and labor-efficient solution. The study uses various metric evaluations for analysis. The Random Forest, XGBoost,K Nearest Neighbour, Adaboost models performs exceptionally great in achieving improved accuracy, enhancing quality assessment processes.

With AdaBoost and other mentioned models, a 98.58% success rate is achieved using 1059 samples. Visual presentations, including confusion matrices, illustrate the comparison of results. The study concludes that machine learning algorithms offer high accuracy in assessing dairy product quality.

Reference

- Milk Source Identification and Milk Quality Estimation Using an Electronic Nose and Machine Learning Techniques by Fanglin Mu , Yu Gu , Jie Zhang and Lei Zhang.

- Eurasian Journal of Food Science and Technology 2022; Vol: 6, Issue: 2, pp:76-87 76
  Using Machine Learning Algorithms to Detect Milk Quality by Ahmet ÇELIK Kütahya Dumlupınar University, Tavşanlı Vocational School, Computer Technologies Department, Kütahya, Türkiye.

- EAI Endorsed Transactions on Internet of Things Research Article
  Milk Quality Prediction Using Machine Learning by Drashti Bhavsar , Yash Jobanputra , Nirmal Keshari Swain , Debabrata Swain.

- Proceedings of the First Australian International Conference on Industrial Engineering and Operations Management, Sydney, Australia, December 20-21, 2022 © IEOM Society International Machine Learning Applied to Milk Sample Classification by Mia León and Diego Ossa, Universidad de Lima, Carrera de Ingeniería Industrial Lima, Perú.

- International Research Journal of Modernization in Engineering Technology and Science
  (Peer-Reviewed, Open Access, Fully Refereed International Journal)
  Volume:05/Issue:05/May-2023 Impact Factor- 7.868 International Research Journal of Modernization in Engineering, Technology and Science

ADULTERATION IDENTIFICATION OF MILK USING MACHINE LEARNING AND IOT

Prof. Ganesh Nikam , Govardhan Gomashe , Abhishek Shinde , Dipali More , Akshata Kave.

Experimental Details

Used kaggle Notebook for all experimentation.

Python packages :

- Sklearn
- Pandas
- Numpy
- Matplotlib
- Seaborn
- Xgboost
- Warnings

NOTEBOOK LINK -
https://www.kaggle.com/code/abhi3022/notebook34395748f7