Support vector machine and decision trees successful in predicting milk quality

Halleigh Kelchen
Department of Engineering, Computing, and Mathematical Sciences
Lewis University
Romeoville, IL, United States
halleighlkelchen@lewisu.edu

***Abstract*—Milk and dairy products are an essential source of nutrients, minerals and amino acids in the human diet. However, microorganisms also find milk to be an excellent habitat for breeding [2]. Using features such as color, fat, turbidity, odor, pH, and temperatures of the milk can be used to determine the quality of milk prior to distribution [3]. In this study, a variety of a machine learning techniques were implemented to predict the quality of milk using Python. Decision trees and support vector machine (SVM) were found to be the most effective supervised machine learning techniques for predicting milk quality.**

Keywords— milk, quality, machine learning, SVM, decision trees

I. Introduction

Milk and dairy products are a leading source of essential micronutrients in the human diet. Milk contains calcium, protein, vitamin D, potassium, and phosphorus, which are important for bone strength and growth [1]. While milk is a nutritious source of nutrients, minerals, amino acids for humans, it can also be a growth substrate for microbes. [2]. Microorganisms can originate anywhere from the milking process to the pasteurization process [2]. This is why the safety and quality of the milk being distributed is of utmost importance [2].

A grading system for milk products is used to classify the milk based on quality and safety standards. A variety of factors, such as color, fat, turbidity, odor, pH, and temperatures are used to group the products [3]. There are three grades that the milk is classified into: low, medium, and high [3].

Previous methods of grading require substantial manual work and time [3]. Supervised machine learning techniques can be implemented to help recognize patterns and classify milk samples by grade in a much more efficient and cost-effective way. Machine learning methods can be used to predict the grade of the milk using color, fat, turbidity, odor, pH, and temperatures.

In this study, Keras and Sckit-Learn were used to apply machine learning methods to a Milk Quality Prediction dataset obtained from Kaggle using Python [4]. A variety of machine learning methods were implemented, such as Gaussian Naïve Bayes, logistic regression, neural network, support vector machine (SVM), and decision tree algorithms. Based on the results, the SVM and the decision tree algorithms generated the most successful predictions.

II. Materials and Methods

A. Data Set and Kaggle

The dataset was obtained from Kaggle, which is an online community that provides access to datasets for the purpose of other scientists and learners to utilize for machine learning and other data science techniques. The Milk Quality Prediction dataset used for this study was taken from Kaggle [4]. There are 7 independent variables: pH, temperature, taste, odor, fat, turbidity, and color [4]. The dependent variable (or target) of this study is the grade, which is classified as either low, medium, or high [4]. Prior to implementing machine learning techniques, the data was preprocessed using sklearn methods such as LabelEncoder and StandardScaler. Grade was changed from a categorical variable to numeric values for easier implementation of machine learning methods. The dataset was then split into train and test data using the train test split feature in sklearn. The test size of this study was 0.25 of the dataset.

B. Gaussien Naïve Bayes

Naïve Bayes uses simple Bayesian networks that separates the unobserved from the observed nodes [10]. This method is based on estimating, which is why some more complex algorithms tend to be more accurate [10]. In this study, Gaussian Naïve Bayes was done using sklearn.

C. Logistic Regression

Logistic regression is a classification technique that uses a single multinomial model with an estimator [10]. This algorithm determines probabilities based on distances from the boundaries between classes [10]. Logistic regression can be used for prediction and is commonly used in statistical applications [10]. In this study, logistic regression was done using sigmoid activation and a L2 learning rate of 0.01. The binary cross entropy vs epoch was measured to determine the model loss.

D. Neural Network

Neural Networks are used to perform both classification and regression techniques. Input and activation functions, architecture, and weights are used to build these networks [10]. In this study, a neural network with four layers was used. Both Relu and sigmoid activations were used, along with the Adam optimizer and a learning rate of 0.01. The binary cross entropy was used to measure the loss of the model.

E. Support Vector Machiene

SVMs, a newer supervised learning technique is like multilayer perceptron neural networks [10]. A margin is used to create the largest possible distance between both sides of the hyperplane [10]. In this study, SVM was implemented using the SVC feature in sklearn.

F. Decision Tree

Decision trees are a classification technique used to sort instances by feature values [10].. This learning method uses a decision tree to map observations and serves a predictive model [10]. The decision tree classifier is sklearn was used in this study.

III. Results and discussion

Several different machine learning algorithms were used to evaluate the performance of the different algorithms. Gaussian Naïve Bayes, logistic regression, neural network, SVM, and decision trees were used to predict the milk quality. The classification models were evaluated using an accuracy measure, confusion matrix, and classification report. The regression methods were evaluated using accuracy measure and binary cross entropy loss.

A. Model Performance

The Gaussian Naïve Bayes algorithm had an accuracy of 0.25. This algorithm had a precision score of 0.33, recall of 0.08, and F1 score of 0.13. The logistic regression had a final loss after iteration of -12.8 and an accuracy of 0.47. Fig. 1 shows the logistic regression model loss of both the training and testing datasets. The neural network had an accuracy of 0.52. This model also had a precision score of 0.33, recall of 0.08, and F1 score of 0.13. The loss after final iteration was -264.32 (Fig. 2.). The SVM had an accuracy score of 0.86. For this model, the precision score was 0.86, recall 0.87, and F1 score of 0.86. The decision tree algorithm had an accuracy of 0.99. The precision, recall, and F1 score for this model were all 1.00.

B. Discussion

Compared to the other models, the SVM and decision tree algorithms had much higher accuracy, precision, recall, and F1 scores. This could be because SVM and decision tree both work well with non-linear solutions, whereas Gaussian Naïve Bayes and logistic regression are both linear classifiers. The results show that seems to be a nonlinear relationship between the variables in the data set and the milk quality. The neural network likely needed more data than the dataset provided to create an accurate model.

C. Limations

There are several limitations of this study. The dataset used in this study contained less than two thousand entries, which is relatively small; a larger data set might generate more accurate results and allow for a better neural network model to be constructed. The data used in this study also does not specify how and where the data was collected from. It is unknown how good this data collection is, although it does not appear to be artificially generated. Depending on these factors, increasing the collection of data could be beneficial in order to create more accurate model predictions.

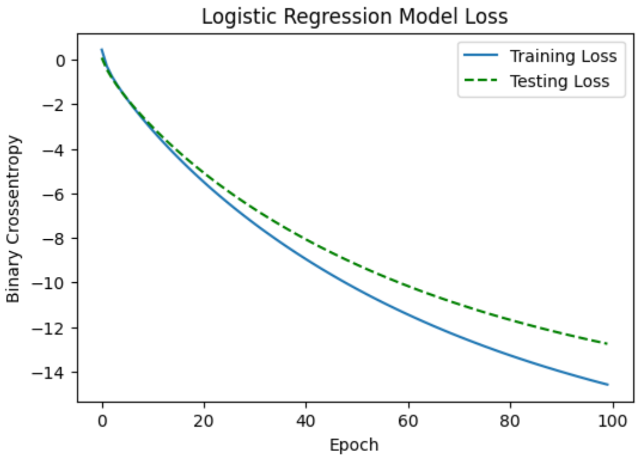

Fig. 1. The model loss from the logistic regression.

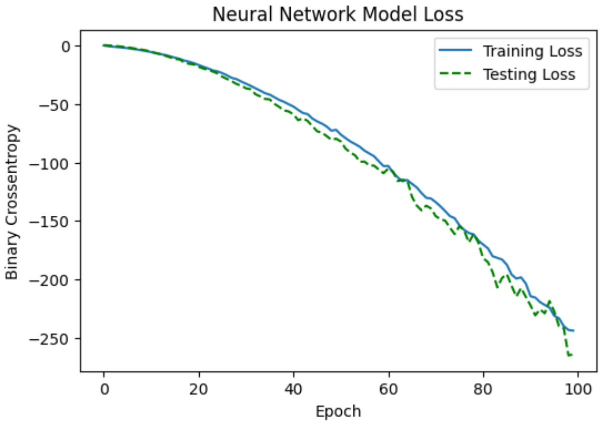

Fig. 2. The model loss from the neural network.

IV. Conclusion

This study showed that milk quality can be predicted using machine learning methods. The best methods for predicting the quality of milk were SVM and decision trees. This study showed that there is a nonlinear relationship between the features and the grade of the milk, making gaussian naïve bayes and logistic regression ineffective for predicting the quality of milk. Increasing the entries in the dataset might allow for a better neural network to be formed. Future research could examine the relationships between the different features of the milk and see which ones are most important in determining the milk quality. Overall, this study was successful in predicting the grade of milk using SVM and decision trees with accuracies over 85 percent.

References

[1] H. Górska-Warsewicz, K. Rejman, W. Laskowski, and M. Czeczotko, “Milk and dairy products and their nutritional contribution to the average Polish

diet," *Nutrients*, vol. 11, no. 8, p. 1771, 2019. doi:10.3390/nu11081771

[2] V. Fusco *et al.*, "Microbial Quality and safety of milk and Milk Products in the 21st Century," *Comprehensive Reviews in Food Science and Food Safety*, vol. 19, no. 4, pp. 2013–2049, 2020. doi:10.1111/1541-4337.12568

[3] W. Habsari, F. Udin, and Y. Arkeman, "An analysis and design of fresh milk smart grading system based on internet of things," *IOP Conference Series: Earth and Environmental Science*, vol. 1063, no. 1, p. 012059, 2022. doi:10.1088/1755-1315/1063/1/012059

[4] S. Rajendran, "Milk Quality Prediction," Kaggle, https://www.kaggle.com/datasets/cpluzshrijayan/milkquality/data (accessed Oct. 22, 2023).

[5] O. F.Y *et al.*, "Supervised machine learning algorithms: Classification and comparison," *International Journal of Computer Trends and Technology*, vol. 48, no. 3, pp. 128–138, 2017. doi:10.14445/22312803/ijctt-v48p126