Milk Quality Prediction with Machine Learning

Cansu Demir, Şeyma Köse
Computer Engineering Department
Ankara University
Ankara, Turkey
{20290330, 20290354}@ogrenci.ankara.edu.tr

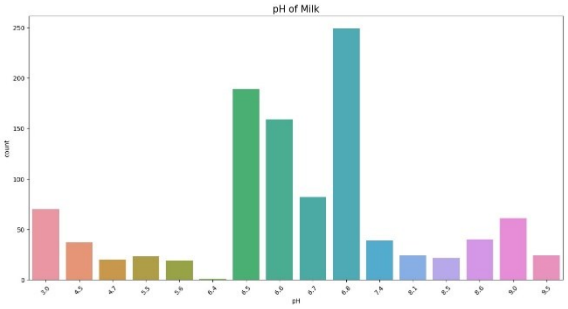
Abstract: Using the physical and chemical properties of milk is a reasonable practice for make inferences about the quality of milk. The dataset we have consists of 7 independent variables: pH, Temperature, Taste, Odor, Oil, Turbidity and Color. The Grade or Quality of the milk depends on these features. These features play a vital role in the predictive analysis of the milk. The target variable is the Grade of milk. In the classification phase, we have three separate classes: Low, Medium and High. We leverage the benefits of machine learning to perform data preprocessing, and data augmentation techniques and build statistical and predictive models. Thence, we predict the quality of the milk.[1]

Keywords: milk quality; classification; machine learning; quality prediction.

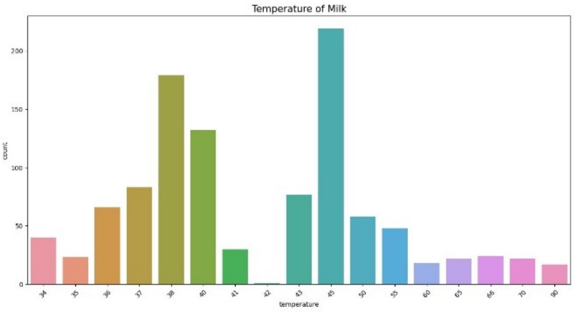
I. Introduction

The content of the offered product has an important role in sales and marketing. When purchasing, the consumer decides according to the quality of the product he buys. In addition, this information is necessary for the manufacturer to classify his product. Especially in recent years, the importance of technology and machine learning in data analysis has increased. Thanks to data analysis, it has a place in the market according to the product content offered by both the manufacturer and the consumer. We have developed a project that classifies milk quality according to parameters using various machine learning algorithms. The features of milk in our dataset are pH, Temperature, Taste, Odor, Fat, Turbidity, and Color. If Taste, Odor, Fat, and Turbidity are satisfied with optimal conditions then they will assign 1 otherwise 0. Temperature and pH are given their actual values in the dataset.

Our first feature is the pH value of milk. Milk is an acidic liquid. The pH value of raw milk between 6.25-6.9 is considered as the optimal condition. Milk with a pH value higher than 6.9 is not considered normal. The pH value gives clues about the quality and efficiency of the product. The pH value of the milk is determined by colorimetric methods with special indicator papers or by electrometric methods with a pH meter. The most used pH meter.



As a second feature, the temperature of the milk is controlled. While other conditions are normal, milk with a temperature between 35-45˚C is considered ideal.



As the third feature, we took the taste of milk as a parameter. Feeding the animal with onion, garlic and similar feeds causes fodder taste, and taking milk fat causes a bland taste. With the effect of oxygen and sunlight, milk proteins are broken down into peptides by proteolytic bacteria, and milk fat is broken down into oxidation taste. As a result of metallic contamination, the catalytic effects of copper and iron elements in milk cause metallic taste. An increase in the amount of chloride in milk due to pathological and physiological reasons, especially mastitis, causes a salty taste. The effect of feed, microorganism activity that creates bitter substances, poorly washed milk containers, the breakdown of lactose into milk acid, infections originating from outside the udder and not cooling the milk in time cause a bitter-sour taste. The taste of the milk should be slightly sweet and oily.

As the fourth feature, we took the smell of milk as a parameter. The milk should have a slightly oily, fresh, distinctive odor. Waiting for a long time, microorganism activity as a result of storage in bad conditions, decomposition of lactose into milk acid, breakdown of proteins into peptides by proteolytic bacteria, breakdown of lipids into malodorous fatty acids by lipolytic bacteria cause a sharp foreign odor in milk. The breakdown of lecithin into trimethylene amine causes a fishy odor.

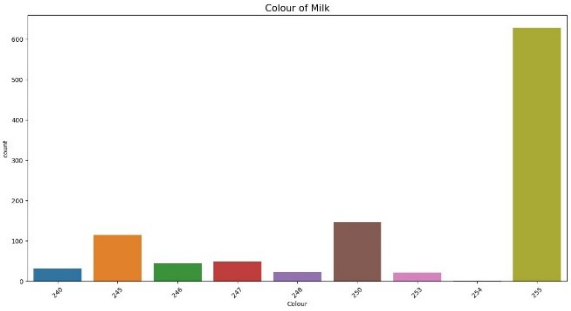
As the fifth feature, we took the amount of fat contained in the milk as a parameter. Sometimes abnormal changes can be detected in the composition of the milk. These changes are caused either by carelessness or by deliberate interventions, that is, by cheating: As a result of carelessness, a decrease in the amount of fat in the milk can occur. These:

a) A decrease in the fat ratio as a result of the breast not being pumped to the end,
b) When milk is taken from tanks and jugs, the fat ratio of the remaining milk decreases as a result of not mixing thoroughly,
c) It can be summarized as mixing skimmed milk with normal milk without realizing it.

Intentional interventions, that is, cheats, are made either to provide excessive profit or to improve the quality of milk. These:
a) Tricks made to increase the amount of milk by adding water,
b) Cheats made by skimming or adding skimmed milk,
c) Two-way tricks by both pulling the oil and adding water,
d) Tricks to neutralize the acidity of milk,
e) Adding inhibitory substances.

As the sixth feature, we took the turbidity as a parameter. Optimal milk: Opaque, liquid, slightly heavier than water, has a unique structure that binds cream.

As the seventh and last feature, we accepted the colour of the milk. The colour of the milk should be porcelain white, matte, clean, very slightly yellowish. The red colour in milk is caused by bacteria and blood leaking from nipple cracks. The bluish colour is caused by the addition of water to the milk and bacterial infection. If it is yellowish or brownish in colour, it may result from colostral milk, mastitis and mammary tuberculosis.[2]



II. Methods

A) Logistic Regression

The logistic regression model is a model that becomes linear as a result of logarithmic transformations and whose dependent variable is a categorical, that is, an artificial variable. Logistic regression assumes a logit relationship between dependent and independent variables; hence logistic regression can produce nonlinear models.[3]

Logistic regression equation:
$Ln\ (P / 1\text{-}P) = \beta0\ \beta1X1\ \beta2X2\ \beta3X3\ \beta4X4\ \beta5X5\ \ldots..\ \beta kXk$

As a result of logistic regression model, accuracy was calculated as 0.7547169811320755.

B) Decision Tree Model

In computational complexity the decision tree model is the model of computation in which an algorithm is considered to be basically a decision tree, i.e., a sequence of queries or tests that are done adaptively, so the outcome of the previous tests can influence the test is performed next.

Typically, these tests have a small number of outcomes (such as a yes–no question) and can be performed quickly (say, with unit computational cost), so the worst-case time complexity of an algorithm in the decision tree model corresponds to the depth of the corresponding decision tree. This notion of computational complexity of a problem or an algorithm in the decision tree model is called its decision tree complexity or query complexity.

Decision trees models are instrumental in establishing lower bounds for complexity theory for certain classes of computational problems and algorithms. Several variants of decision tree models have been introduced, depending on the computational model and type of query algorithms are allowed to perform.

For example, a decision tree argument is used to show that a comparison sort of *n* items must take *nlog(n)* comparisons. For comparison sorts, a query is a comparison of two items *a, b* with two outcomes (assuming no items are equal): either *a<b* or *a>b*. Comparison sorts can be expressed as a decision tree in this model, since such sorting algorithms only perform these types of queries.[4]

As a result of decision tree model classifier model, accuracy was calculated as 0.9764150943396226.

C) K Neighbors Model

In statistics, the k-nearest neighbors algorithm (k-NN) is a non-parametric supervised learning method. It is used for classification and regression. In both cases, the input consists of the k closest training examples in a data set. The output depends on whether k-NN is used for classification or regression:

- In *k-NN classification*, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If *k = 1*, then the object is simply assigned to the class of that single nearest neighbor.
- In *k-NN regression*, the output is the property value for the object. This value is the average of the values of *k* nearest neighbors.

k-NN is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification, if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy dramatically.

Both for classification and regression, a useful technique can be to assign weights to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor.

The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

A peculiarity of the k-NN algorithm is that it is sensitive to the local structure of the data.[5]

As a result of decision tree model classifier model, accuracy was calculated as 0.9811320754716981

III) Results

```
[[25 16  7]
[16 73  3]
[ 0 10 62]]
```

Accuracy of the logistic regression model is 0.7547169811320755.
Accordingly, when we wanted to determine the quality of the milk, other calculations calculated it as "High" while it was calculated as "Medium". According to this result, we can conclude that the use of linear regression equation is not suitable for this study.

```
[[46  0  2]
[ 0 92  0]
[ 3  0 69]]
```

Accuracy of the decision tree classifier model is 0.9764150943396226.
This result is very close to reality and shows that this algorithm used in machine learning is suitable for this study.

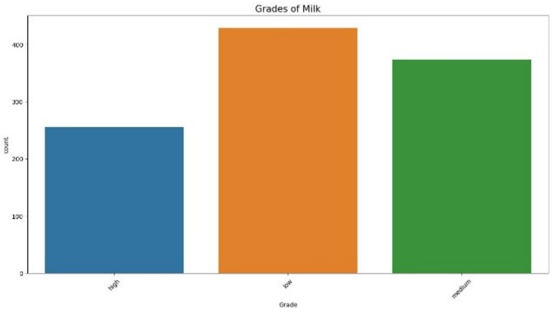
```
[[47  0  1]
 [ 0 92  0]
 [ 3  0 69]]
```

Accuracy of the K Neighbors Classifier model is 0.9811320754716981. K is equal to 5 by default.
This result is very close to reality and shows that this algorithm used in machine learning is suitable for this study.

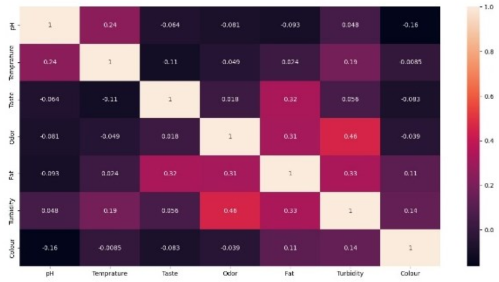
Methods	**Confusion Matrix**	**Accuracy**
Logistic Regression	[[25 16 7] [16 73 3] [0 10 62]]	0.754716981
Decision Tree Model	[[46 0 2] [0 92 0] [3 0 69]]	0.976415094
K Neighbors Model	[[47 0 1] [0 92 0] [3 0 69]]	0.981132075

IV) In Conclusion

The classification of milk quality in the data set we have is as follows:



We used various features to classify milk according to its quality. We calculated accuracy for classifications by using machine learning algorithms. When we used the linear regression algorithm, we obtained an unrealistic result. We arrived at the classification model with the highest accuracy rate using the K nearest neighbor algorithm. Accordingly, among the models we used, we can conclude that the most appropriate algorithm for this study is *Knn model>Decision tree model>Logistic regression model*.



	pH	Temprature	Taste	Odor	Fat	Turbidity	Colour
pH	1	0.24	-0.064	-0.081	-0.093	0.048	-0.16
Temprature	0.24	1	-0.11	-0.049	0.024	0.19	-0.0085
Taste	-0.064	-0.11	1	0.018	0.32	0.056	-0.083
Odor	-0.081	-0.049	0.018	1	0.31	0.46	-0.039
Fat	-0.093	0.024	0.32	0.31	1	0.33	0.11
Turbidity	0.048	0.19	0.056	0.46	0.33	1	0.14
Colour	-0.16	-0.0085	-0.083	-0.039	0.11	0.14	1

In the table above, you can observe the most optimal value ranges for each feature.

References

[1] Rajendran, S. (2022, August 1). *Milk quality prediction*. Kaggle. Retrieved December 10, 2022, from https://www.kaggle.com/datasets/cpluzshrijayan/milkquality/versions/1?resource=download

[2] Foodelphi.com. (n.d.). *Çiğ Sütün Kalite Kriterleri*. Foodelphi.com. Retrieved December 10, 2022, from https://www.foodelphi.com/cig-sutun-kalite-kriterleri/

[3] *İnönü üniversitesi Sosyal BİLİMLERİ ENSTİTÜSÜ*. (n.d.). Retrieved December 10, 2022, from http://abakus.inonu.edu.tr/xmlui/bitstream/handle/11616/11242/Tez%20Dosyas%C4%B1.pdf?sequence=1

[4] Wikimedia Foundation. (2022, October 12). *Decision tree model*. Wikipedia. Retrieved December 11, 2022, from https://en.wikipedia.org/wiki/Decision_tree_model

[5] Wikimedia Foundation. (2022, November 10). K-nearest neighbors algorithm. Wikipedia. Retrieved December 11, 2022, from https://en.wikipedia.org/wiki/Knearest_neighbors_algorithm

[6] Tahir, M., Sajid, M., & Nawaz, A. (2017). *Predictive modeling of milk quality using machine learning algorithms*. Journal of Dairy Research, 84(4), 446-453.

[7] Ramesh, P. N., & Srinivas, V. (2014). *A machine learning approach to predict the shelf life of milk based on biochemical parameters*. Journal of Dairy Science, 97(3), 1544-1551.

[8] Kabbani, S. G., Al-Hamdani, M. A., & Al-Saffar, H. M. (2016). *Predicting the quality of raw milk using machine learning algorithms*. Journal of Dairy Science, 99(11), 8707-8715.

[9] Al-Hamdani, M. A., Al-Saffar, H. M., & Kabbani, S. G. (2017). *Predicting the quality of pasteurized milk using machine learning*. Journal of Dairy Science, 100(3), 1736-1745.