International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 04 | April - 2023 Impact Factor: 8.176 ISSN: 2582-3930

MILK QUALITY PREDICTION USING RAPID MINER AND ML ALGORITHMS

Dr.V. Savithri[1], Jothi Bathra L[2], Abarna K[2], Dharani P[2]

[1] Assistant professor, Department of Computing, Coimbatore Institute of Technology, India

[2] Student, Department of Computing – Decision and Computing Sciences, Coimbatore Institute of Technology, India

***Abstract*— Milk is an essential source of nutrition for humans, and its quality is crucial for ensuring food safety and public health. The quality of milk is dependent on various factors, such as animal health, feeding practices, and environmental conditions. Developing a predictive model to assess the quality of milk using machine learning techniques. This study aims to investigate the potential of using RapidMiner and R language to predict the quality of milk using a random forest algorithm. The dataset used in this study includes various parameters related to milk quality, such as fat content, protein content, lactose, and somatic cell count. The data is preprocessed and then fed into the random forest algorithm for training and testing the model. The results of the study demonstrate that the developed model is capable of accurately predicting milk quality, which can be beneficial for the dairy industry to ensure milk quality and enhance the product's overall value.**

I. INTRODUCTION

Milk is an essential source of nutrition for humans, and its quality is crucial for ensuring food safety and public health. The quality of milk is dependent on various factors, such as animal health, feeding practices, and environmental conditions. It is challenging to assess milk quality accurately using conventional methods, and there is a need to develop reliable and efficient techniques for this purpose. Machine learning techniques have shown great potential for predicting the quality of milk.

The aim of this study is to develop a predictive model for assessing the quality of milk using machine learning techniques. Specifically, employ RapidMiner and R language to build a random forest model for predicting milk quality. The dataset used in this study includes various parameters related to milk quality, such as fat content, protein content, lactose, and somatic cell count. The data is preprocessed and then fed into the random forest algorithm for training and testing the model.

The study aims to demonstrate the potential of machine learning techniques for predicting milk quality accurately. The developed model can be useful for the dairy industry to ensure milk quality and enhance the product's overall value. Additionally, the study can help in identifying the critical factors that affect milk quality, which can aid in developing strategies for improving milk production and quality. Overall, this study can contribute to ensuring the safety and quality of milk products and their impact on public health.

II. METHODOLOGY

A. DATASET DESCRIPTION

This dataset is manually collected from observations. It helps to build machine learning models to predict the quality of milk. This dataset consists of 7 independent variables ie pH, Temperature, Taste, Odor, Fat, Turbidity, and Color. Generally, the Grade or Quality of the milk depends on these parameters. These parameters play a vital role in the predictive analysis of the milk. The target variable is nothing but the Grade of the milk. 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

Ph: This Column defines PH values of the milk which ranges from 3 to 9.5 max : 6.25 to 6.90

Temperatures: Temperatures of the milk which range from 34'C to 90'C max : 34'C to 45.20'C.

Taste: Taste of the milk which is categorical data 0 (Bad) or 1 (Good) max : 1 (Good)

Odor: Odor of the milk which is categorical data 0 (Bad) or 1 (Good) max : 0 (Bad)

Fats: Fat of the milk which is categorical data 0 (Low) or 1 (High) max : 1 (High)

Turbidity: Turbidity of the milk which is categorical data 0 (Low) or 1 (High) max : 1 (High)

Colour: Colour of the milk which ranges from 240 to 255 max : 255

Grades: (Target) of the milk which is categorical data Where Low (Bad) or Medium (Moderate) High

	pH	Temprature	Taste	Odor	Fat	Turbidity	Colour	Grade
0	6.6	35	1	0	1	0	254	0
1	6.6	36	0	1	0	1	253	0
2	8.5	70	1	1	1	1	246	1
3	9.5	34	1	1	0	1	255	1
4	6.6	37	0	0	0	0	255	2
...	...	...	...	...	...	...	...	...

Figure 2.1 Dataset description

B. PROCESSING STEP

Data cleaning is an essential step in preparing the data for machine learning modeling. The following are some steps for data cleaning in milk quality prediction:

1. *Handling missing data:* If there are any missing values in the dataset, they need to be handled appropriately. One common technique is to replace missing values with the mean or median of the feature. Another approach is to remove the samples with missing values.

2. *Handling outliers:* Outliers are data points that lie far from the majority of the data points and can adversely affect the model's performance. They need to be identified and handled appropriately. One common technique is to remove the outliers or transform the data using techniques such as log transformation.

3. *Handling categorical data:* If the dataset contains categorical features, they need to be converted into numerical data that can be used in machine learning models. One approach is to use one-hot encoding, where each category is converted into a binary feature.

4. *Scaling data:* Scaling is essential to ensure that the features are on a similar scale. Common techniques include min-max scaling or standard scaling.

5. *Removing redundant features*: If the dataset contains redundant or highly correlated features, they can be removed to reduce the model's complexity and improve its performance

C. MODEL BUILDING STAGE

After performing data cleaning, the next step is to build the machine learning model for milk quality prediction. In this case, random forest algorithm is used to build the model. The following are the steps for building the model:

1. *Split the dataset into training and testing sets:* The dataset needs to be divided into two parts: one for training the model and the other for testing the model's performance. Typically, 70-80% of the data is used for training, and the rest is used for testing.

2. *Feature selection*: It is essential to select the most relevant features for the model to reduce the dimensionality of the data and improve the model's performance. This can be done using techniques such as correlation analysis or feature importance analysis.

3. *Building the model:* After feature selection, build the random forest model using the training data. The model will learn to predict milk quality based on the selected features.

4. *Testing the model:* Finally, test the model's performance using the testing dataset. Evaluate the model's performance using metrics such as accuracy, Confusion matrix, precision, recall, and F1 score.

By following these steps,build an accurate and reliable random forest model for predicting milk quality.

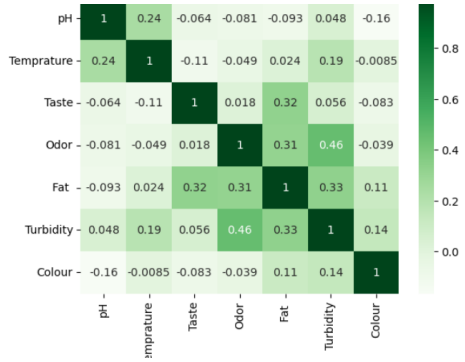

Figure 2.2 Correlation map among the variables

```
Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            0.9873   1.0000   1.0000
Specificity            1.0000   0.9948   1.0000
Pos Pred Value         1.0000   0.9896   1.0000
Neg Pred Value         0.9952   1.0000   1.0000
Prevalence             0.2762   0.3322   0.3916
Detection Rate         0.2727   0.3322   0.3916
Detection Prevalence   0.2727   0.3357   0.3916
Balanced Accuracy      0.9937   0.9974   1.0000
```

Figure 2.3 Summary statistics

III. ALGORITHMS USED

A. RANDOM FOREST CLASSIFIER

Random forest classifier is a machine learning algorithm that can be used for milk quality prediction. It is a type of ensemble learning method that uses multiple decision trees to make predictions. Each decision tree is trained on a random subset of the data, and the final prediction is based on the majority vote of all the decision trees. Overall, a random forest classifier can be a useful tool for milk quality

prediction, as it can handle complex datasets with many features and can provide accurate predictions with relatively low computational cost.

B. DECISION TREE CLASSIFIER

A decision tree classifier is a machine learning algorithm that can be used for milk quality prediction. It works by recursively splitting the data into smaller subsets based on the values of different features, until the subsets are pure or nearly pure in terms of their target variable. One advantage of decision tree classifiers is that they are easy to interpret, as the resulting tree can be visualized and used to identify the most important features for predicting milk quality. However, decision trees can also be prone to overfitting if the tree is too deep or if there are too many features, so it is important to tune the hyperparameters carefully to avoid overfitting.

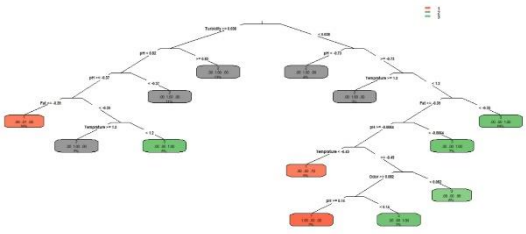

Figure 3.1 Decision Tree

IV. RAPID MINER

RapidMiner is a data science platform that provides a range of tools and operators for predictive modeling, including data preparation, modeling, validation, and deployment. It allows users to build and deploy predictive models without writing any code, using a visual workflow interface.

1. Milk quality dataset into RapidMiner separate the features (e.g., fat content, protein content) and the target variable (e.g., milk quality label) into different columns.

2. Preprocess the dataset to handle missing values, outliers, and other issues RapidMiner's data preparation operators to perform tasks such as imputation, normalization, and feature selection.

3. Split dataset into training and testing sets using rapidminer's data splitting operators. Evaluate the performance of milk quality prediction model on unseen data.

4. Select a machine learning algorithm to use for milk quality prediction. RapidMiner provides a range of classification algorithms, including decision trees, random forests, and support vector machines. Compare the performance of different algorithms using RapidMiner's validation operators.

5. Train milk quality prediction model on the training set using RapidMiner's modeling operators. This will involve specifying the algorithm and its parameters, as well as the features and target variable to use. Here regression classifier and random forest classifier are used.

6. Evaluate the performance of milk quality prediction model on the testing set using RapidMiner's performance operators. Metrics such as accuracy, precision, recall, and F1 score, which you can use to assess the quality of model.

7. Milk quality prediction model to make predictions on new data. RapidMiner's scoring operators to apply the model to new datasets and generate predictions for the milk quality label.

Overall, RapidMiner provides a range of tools and operators for performing milk quality prediction using machine learning. By following these steps and experimenting with different algorithms and parameters, you can create a robust and accurate model for predicting the quality of milk.

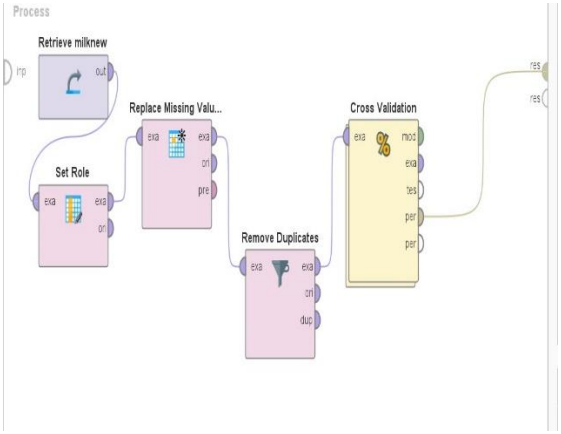

Figure 4.1 Rapid miner Process

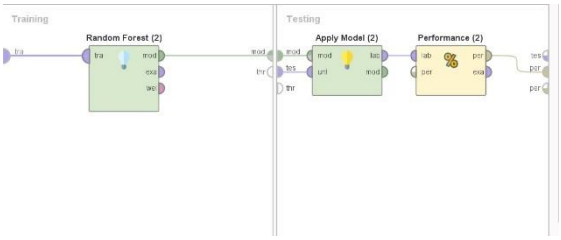

Figure 4.2 Rapid miner Process

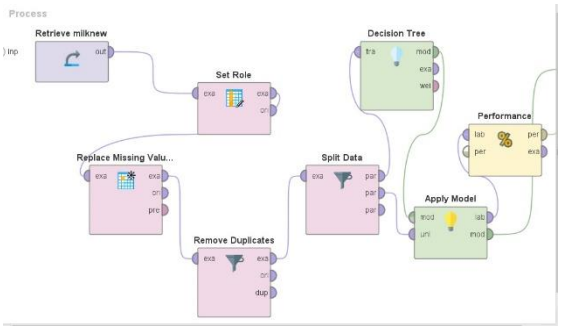

Figure 4.3 Rapid miner Process

V. RESULT ANALYSIS

Accuracy is the percentage of correctly classified instances out of the total instances in the test set. Precision is the percentage of correctly classified positive instances (i.e., good quality milk) out of all instances classified as positive. Recall is the percentage of correctly classified positive instances out of all actual positive instances. F1 score is the harmonic mean of precision and recall. Overall, analyzing the results of milk quality prediction using a decision tree classifier in RapidMiner involves using performance metrics to evaluate the model's accuracy and effectiveness, identifying patterns and trends in the data that may be affecting the model's performance, and experimenting with different hyperparameters to optimize the model's performance.

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3
         1  78   0   0
         2   1  95   0
         3   0   0 112

Overall Statistics

               Accuracy : 0.9965
                 95% CI : (0.9807, 0.9999)
    No Information Rate : 0.3916
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9947

 Mcnemar's Test P-Value : NA
```

Figure 5.1 Confusion Matrix

VI. CONCLUSION

In conclusion, milk quality prediction using machine learning (ML) can be a useful tool in the dairy industry to improve the quality of milk products and increase consumer satisfaction. By using ML algorithms such as decision trees, random forests.it is possible to accurately predict the quality of milk based on its various features.

Overall, ML-based milk quality prediction has the potential to greatly benefit the dairy industry by allowing for more efficient and accurate quality control processes, leading to higher quality milk products and greater customer satisfaction.

VII. REFERENCES

[1] "Machine Learning: A Probabilistic Perspective" by Kevin Murphy

[2] "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" by Aurélien Géron

[3] Franken, H. and Körner, A. (2021). Predictive Modeling with RapidMiner Studio. RapidMiner. Retrieved April 12, 2023, from https://rapidminer.com/predictive-modeling-with-rapidminer-studio/