**MILK QUALITY PREDICTION**

# PROBLEM STATEMENT

Milk is one the highly consumable food product with high nutritious value across the globe. But it gets spoiled easily as well because of contamination with the bacteria. It needs to be handled carefully. Consuming bad quality milk can lead to several health issues. Determining the quality of milk is now a very important aspect for both the dairy industry as well as the consumers. Quality of milk can be determined based on various factors like the odour of the milk, colour of the milk, taste, the temperature, PH, etc. This study intends to create a solution in the form of ml model that can help to predict the quality of the milk based on these factors.

# METHODOLOGY

To build this model multiple supervised machine learning algorithms will be trained and tested. The model with best performance and accuracy will be considered ideal for the prediction of the quality of the milk in new test samples based on the features of the milk.

For building the solution, first the dataset was loaded and pre-processed. In the dataset, no missing values were found. The dataset has the shape of 1059 rows and 8 columns. The dependent column called as grade has three categories defining the quality of milk, which is low, medium and high.

Exploratory data analysis was done, and multiple insights were found. Some of them were –

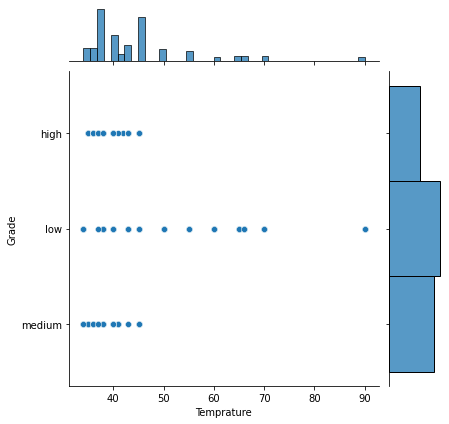
* PH of the milk was analysed and compared over different qualities of milk. It was found that usually low-quality milk have very high and very low PH values.

A picture containing screenshot, rectangle, diagram, square

Description automatically generated

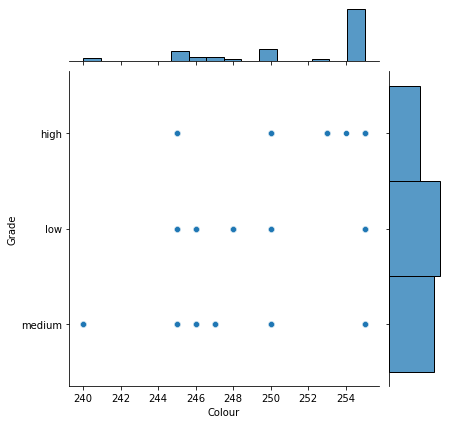
**Figure 1 EDA1**

* Temperature of milk was analysed as per quality of milk. Low quality milk seems to have different temperatures.



**Figure 2 EDA2**

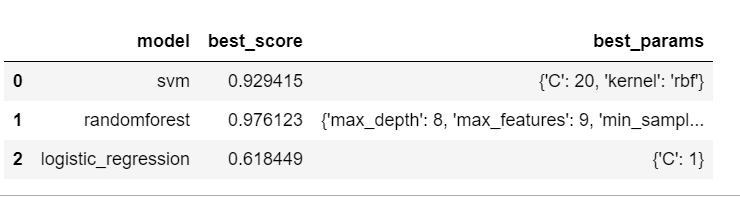
* Colour of the milk seems to have no impact on the quality of milk.



**Figure 3 EDA3**

Label encoding was done to encode all the object datatype features into numerical one. Correlation matrix was made to analyse multicollinearity in the dataset. No multicollinearity in the dataset was found. Standardization of the data was done over the dataset. Principal component analysis was done over the data and two principal components were made By reducing a large collection of variables into a smaller one that still maintains much of the information in the large set, principal component analysis (PCA) is a dimensionality reduction method frequently used to reduce the dimensionality of huge data sets. Hyperparameter tuning was performed over three models, Logistic Regression, support vector machine, and Random Forest classifier. Grid search cross validation was performed to select best hyper parameters of all three models.

Best parameters and best scores of all three models were as follow -

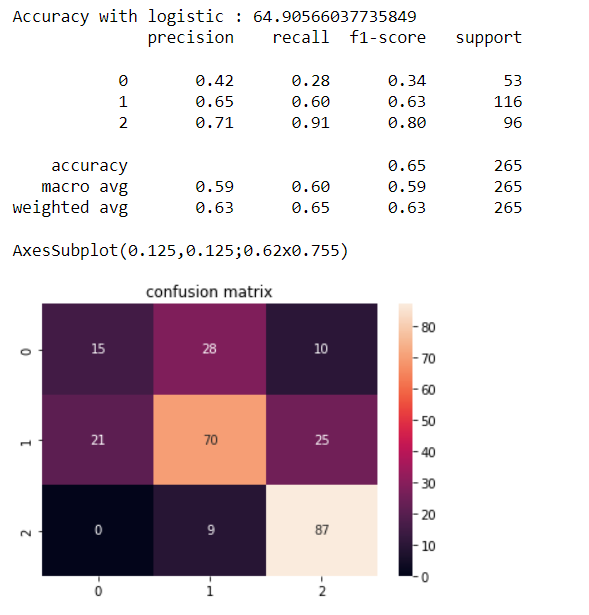


**Figure 4 Results of hyperparameter tuning**

Based on the best parameters all three models were finally trained and evaluated.

# Results

* **Logistic Regression**



**Figure 5 Results of Logistic Regression**

* Random Forest Classifier

A screenshot of a computer

Description automatically generated with low confidence

**Figure 6 Results of Random Forest**

* Support vector classifier

A screenshot of a computer

Description automatically generated with low confidence

**Figure 7 Result of Support vector Classifier**

Random Forest classifier tends to perform best with 98% accuracy.