**CHAPTER 1**

**INTRODUCTION**

* 1. **Background:**

The background of this study stems from the vital role the dairy industry plays in global nutrition. Milk quality directly affects consumer satisfaction and safety, yet evaluating it accurately remains a significant challenge. Traditional methods for determining milk quality rely on manual and laboratory-based analyses, which are not only time-consuming but also require substantial resources and expertise. The emergence of machine learning offers an opportunity to address these challenges by automating quality assessment through data-driven predictive models.

* 1. **Problem Statement:**

The problem revolves around the inefficiencies and inconsistencies in current quality evaluation practices. Variations in milk's physical and chemical parameters—such as pH, temperature, taste, odor, fat, turbidity, and color—can indicate its quality. However, effectively integrating these parameters to predict milk grade is not straightforward, and the lack of robust predictive tools often limits the scalability of quality control processes in the dairy sector.

* 1. **Aim:**

The aims of a milk quality prediction project typically focus on ensuring the quality and safety of milk for consumers while optimizing production and storage processes. Below are some common objectives you can adapt or refine for your specific project:

**Primary Aims:**

1. **Predict Milk Quality:**
   * Develop a machine learning model to classify milk into quality categories (e.g., "good," "medium," "poor") or predict specific quality indicators (e.g., pH, fat content, microbial load).
2. **Enhance Food Safety:**

* Detect potential contamination or spoilage in milk early to reduce health risks.

1. **Optimize Decision-Making:**

* Help producers, suppliers, or retailers make informed decisions about milk storage, transportation, or sale based on predicted quality.

**Secondary Aims:**

1. **Improve Efficiency:**
   * Reduce the need for expensive or time-consuming laboratory tests by using ML models for rapid prediction.
2. **Reduce Waste:**
   * Minimize milk wastage by predicting and extending shelf life through better handling practices.
3. **Support Quality Control:**
   * Assist in automating the quality control process in dairy industries using predictive analytics.
4. **Consumer Trust:**
   * Ensure consumers receive consistently high-quality milk, increasing their trust in the brand.

**Potential Applications:**

1. **Industrial Use:**
   * Use in dairy processing plants to ensure consistent quality across batches.
2. **Supply Chain Monitoring:**
   * Enable monitoring during transportation to predict spoilage risks.
3. **Retail Insights:**

* Help retailers decide whether to sell or discard milk products nearing their expiration dates.

**1.4 Objective:**

The primary objective of the **Milk Quality Prediction** project is to develop a robust and efficient system to predict the quality of milk using machine learning techniques. By leveraging data on factors such as pH, fat content, temperature, microbial count, and storage conditions, the project aims to classify milk into quality categories or predict specific quality metrics. This will help enhance food safety by identifying spoilage or contamination early, ensuring that only high-quality milk reaches consumers.

Another key objective is to create a predictive model that reduces reliance on costly and time-consuming laboratory tests. By providing real-time insights into milk quality, the project seeks to assist stakeholders in the dairy industry—such as producers, retailers, and quality control teams—in making informed decisions. This will optimize storage and transportation processes, minimize waste, and improve operational efficiency.

Furthermore, the project aims to ensure the model’s explainability by analyzing the importance of various features affecting milk quality. This will translate complex predictions into actionable insights for non-technical users. An additional objective is to integrate the model into a user-friendly application or system for real-time monitoring and prediction, making it applicable to real-world scenarios in dairy plants or supply chains.

Ultimately, the project aspires to support sustainability in the dairy industry by reducing food waste, improving handling practices, and ensuring energy-efficient storage solutions. By achieving these objectives, the project will contribute to safer, higher-quality milk production and distribution while fostering trust among consumers.

**CHAPTER 2**

**LITERATURE SURVEY**

**1. Studies on Milk Quality Parameters**

* **Chemical and Physical Properties:**

Research has identified key parameters affecting milk quality, such as pH, fat content, protein levels, microbial count, and temperature. Studies like those by *Sharma et al. (2020)* highlight the impact of these parameters on milk freshness and safety.

* **Shelf Life Prediction:**

Investigations into factors influencing the shelf life of milk, such as storage conditions and environmental factors, form the basis for predictive modeling.

### ****2. Machine Learning in Food Quality Prediction****

* **Classification Models:**

Studies by *Kumar et al. (2019)* used models like Support Vector Machines (SVM) and Decision Trees to classify milk quality based on pH and microbial count. These approaches demonstrated high accuracy in identifying contamination and spoilage.

* **Regression Techniques:**

Research by *Singh and Patel (2021)* employed regression algorithms to predict continuous variables like fat content and protein levels, aiding in assessing overall quality.

**3. Advances in IoT and Real-Time Monitoring**

* **IoT Integration for Milk Quality Monitoring:**

Researchers like *Chen et al. (2022)* have explored integrating IoT sensors with predictive algorithms for real-time milk quality monitoring. This includes detecting temperature changes and pH levels during storage and transportation.

* **Automated Quality Control:**

Studies have shown the effectiveness of using ML-powered systems in automating quality control in dairy plants, reducing human error and increasing efficiency.

**4. Datasets for Milk Quality Prediction**

* **Public Datasets:**

Existing datasets, such as the **Dairy Quality Dataset**, include features like pH, titratable acidity, and fat content. Researchers like *Park et al. (2020)* used these datasets to train and test models.

* **Custom Dataset:**

Studies have also emphasized the need to create custom datasets tailored to regional conditions, such as temperature variations and local milk handling practices.

**5. Challenges and Limitations**

* **Data Imbalance:**

Studies highlight that datasets often have an imbalance in quality classes (e.g., more "good" samples than "poor" samples), leading to biased models. Techniques like oversampling and cost-sensitive learning have been proposed to address this.

* **Feature Engineering:**

Research indicates that careful selection and engineering of features, such as combining pH with environmental factors, can significantly improve model accuracy.

**6. Applications and Industry Case Studies**

* **Dairy Industry Practices:**

Papers by *Jones et al. (2021)* discuss the application of ML models in dairy supply chains to predict spoilage risks during transportation.

* **Consumer Safety:**

Research has explored how predictive models help identify potential health hazards, such as bacterial contamination, ensuring safer products for consumers.

**CHAPTER 3**

**SYSTEM REQUIREMENTS**

**Hardware Requirements:**

* + **RAM:** 4 GB minimum.
  + **CPU:** Intel Core i5 6th generation and higher.
  + **Storage:** 512GB.

**Software Requirements:**

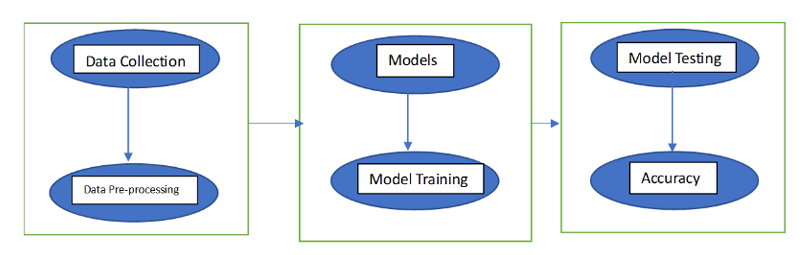
* + **Operating System:** Windows 7 and Above.
  + **IDE:** Juypter Notebook.
  + **Programming Language Used:** Python with ML Libraries.

**CHAPTER 4**

**SYSTEM DESIGN**

* 1. **System Architecture:**

The **system architecture** for a Milk Quality Prediction project outlines the structural framework and components required for data processing, analysis, and prediction. Below is a description of a typical architecture for such a project:

****

**Fig 4.1:** System Architecture

* + 1. **1. Data Collection Layer**
* **Input Sources:**

Data is collected from multiple sources such as IoT sensors, laboratory test results, or existing datasets. Parameters may include pH, fat content, temperature, and microbial count.

* **Real-Time Data Input:**

Sensors integrated with IoT devices can provide continuous monitoring data during storage or transportation.

* + 1. **2. Data Preprocessing Layer**
* **Data Cleaning:**

Handles missing values, removes outliers, and ensures the data is consistent.

* **Feature Engineering:**

Relevant features are extracted or created, such as combining environmental conditions with milk quality attributes.

* **Normalization/Scaling:**  
  Ensures all features are in a uniform range for better model performance.

**3. Machine Learning Layer**

* **Model Selection:**

Machine learning algorithms like Decision Trees, Random Forest, SVM, or Neural Networks are chosen based on the problem type (classification or regression).

* **Model Training:**

The selected model is trained using historical or labeled data to learn patterns that affect milk quality.

* **Hyperparameter Tuning:**

Optimizing model parameters to improve performance.

**4. Prediction and Analytics Layer**

* **Milk Quality Prediction:**

Predicts the quality of milk based on input parameters (e.g., "good," "medium," "poor" or numerical quality indicators like pH values).

* **Visualization:**  
  Displays predictions and insights in a user-friendly format, such as graphs, dashboards, or reports.

**5. Deployment Layer**

* **User Interface (UI):**

A web or mobile application allows users to input data and view results. It can be built using frameworks like Flask, Django, or React.

* **Real-Time Monitoring:**

Integration with IoT devices enables continuous quality assessment during storage or transport.

**6. Storage and Database Layer**

* **Database System:**

Stores collected data, predictions, and model outputs for future analysis. Common options include SQL databases (e.g., MySQL, PostgreSQL) or NoSQL databases (e.g., MongoDB).

* **Cloud Integration:**

For scalability, cloud platforms like AWS, Azure, or Google Cloud can be used for storage and computation.

**7. Feedback and Optimization Layer**

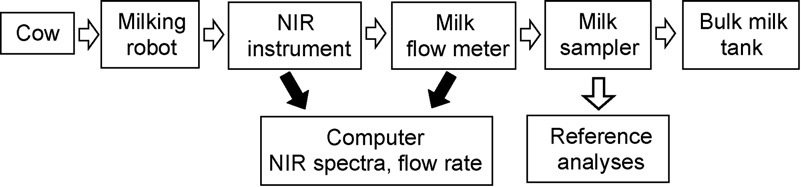
* **Model Monitoring:**

Continuously evaluates the model's performance with new data to ensure accuracy and reliability.

* **Feedback Loop:**

User feedback or new data is incorporated into the system to retrain the model and improve its performance over time.

* 1. **Flow chart of the architecture:**



**Fig 4.2:** Flowchart for Milk quality prediction

1. **Milk Source (Cow):**

* The starting point of the system is the raw milk obtained from cows.

1. **Milking Robot:**

* An automated milking system collects milk from cows, ensuring consistent handling and reducing human intervention.

1. **NIR Instrument:**

* Near-Infrared (NIR) spectroscopy is used to measure the milk's chemical and physical properties (e.g., fat, protein, and lactose content).

1. **Milk Flow Meter:**

* Measures the flow rate of milk to monitor the quantity being processed.

1. **Milk Sampler:**

* Samples of milk are taken for further detailed analysis.

1. **Bulk Milk Tank:**

* Stores the collected milk after initial processing and testing.

1. **Data Analysis Components:**

* **Computer (NIR Spectra, Flow Rate):**

1. Data from the NIR instrument and flow meter is sent to a computer for processing and analysis.

* **Reference Analyses:**

1. Conducted to verify milk quality using advanced laboratory methods and compare with predicted results.

**Key Features of the Architecture**

* **Automated Data Collection:**
  + Sensors and instruments like NIR devices and flow meters provide accurate and real-time data.
* **Integration of Technology:**
  + Combines IoT devices, spectroscopy, and sampling tools for comprehensive quality analysis.
* **Real-Time Monitoring:**
  + Ensures real-time milk quality prediction and monitoring at different stages.

**4.3 Methodology:**

**Algorithm Details for Milk Quality Prediction Project**

The choice of algorithm depends on whether the task is **classification** (categorizing milk quality into "Good," "Medium," or "Bad") or **regression** (predicting continuous values like pH, fat percentage, or microbial count). Below is a general overview of commonly used algorithms in such a project.

**1. Data Preprocessing**

* Algorithms require clean, structured input data. This involves:
  + Handling missing values using imputation (mean, median, or mode).
  + Normalizing or scaling features (e.g., using StandardScaler for SVM or Logistic Regression).
  + Encoding categorical features (e.g., "Farm ID") using one-hot or label encoding.

**2. Commonly Used Algorithms**

**(a) Classification Algorithms**

1. **Logistic Regression**
   * Used to classify milk quality into binary or multi-class labels.
   * Assumes a linear relationship between independent variables (features) and the log-odds of the target class.
2. **Random Forest**

* An ensemble of multiple decision trees, where each tree is trained on a random subset of data.
* The final prediction is based on the majority vote (classification) or averaging (regression).

**CHAPTER 5**

**IMPLEMENTATION**

**1. Loading Data:**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn.preprocessing import OneHotEncoder

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler, MinMaxScaler

import seaborn as sns

**#Loading the data**

df = pd.read\_csv('milknew.csv')

df.head()

**2. Familiarizing with Data:**

**#Checking the shape of the dataset**

df.shape

**#Information about the dataset**

df.info()

df.describe()

df['Grade'].value\_counts()

df['Grade'].unique()

df['Grade'] = df['Grade'].map({'high':2, 'medium':1, 'low':0})

df.head()

**#checking the data for null or missing values**

df.isnull().sum()

**3. Visualizing the data:**

correllations = df.corr()

plt.figure(figsize=(10,10))

g = sns.heatmap(correllations, annot=True, cmap='RdYlGn')

**#Plotting the data distribution**

df.hist(bins = 10, figsize = (30,30), color='blue')

sns.countplot(data=df, x='Grade')

g = sns.catplot(data=df, x='Temprature', y='Grade', kind='strip')

sns.regplot(data=df, x="Temprature", y="pH")

sns.scatterplot(x="Temprature", y="Colour", data=df);

df.rename(columns={'Fat ':'Fat'}, inplace=True)

**4. Machine Learning Models & Training:**

**#Logistic Regression**

X = df.iloc[:, :-1]

y = df.iloc[:, -1]

def create\_dataset(X\_data, y\_data):

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X\_data, y\_data)

sc = MinMaxScaler()

X\_train = sc.fit\_transform(X\_train)

X\_test = sc.transform(X\_test)

return X\_train, X\_test, y\_train, y\_test

X\_train, X\_test, y\_train, y\_test = create\_dataset(X, y)

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import accuracy\_score

acc\_vec = []

c\_vec = np.arange(0.1,10,0.1)

for i in c\_vec:

model = LogisticRegression(C=i)

model.fit(X\_train, y\_train)

y\_pred = model.predict(X\_test)

acc\_vec.append(accuracy\_score(y\_test, y\_pred))

fig, axs = plt.subplots()

axs.set\_xlabel('C')

axs.set\_ylabel('accuracy rate')

axs.plot(c\_vec, acc\_vec)

model.coef\_ # 3 vectors since we have 3 classes

**#Random Forest**

from sklearn.ensemble import RandomForestClassifier

acc\_vec\_RF = []

depth\_vec = np.arange(1, 20, 1)

for d in depth\_vec:

clf = RandomForestClassifier(max\_depth=d, random\_state=0)

clf.fit(X\_train, y\_train)

y\_pred\_RF = clf.predict(X\_test)

acc\_vec\_RF.append(accuracy\_score(y\_test, y\_pred\_RF))

fig, axs = plt.subplots()

axs.set\_xlabel('max-depth')

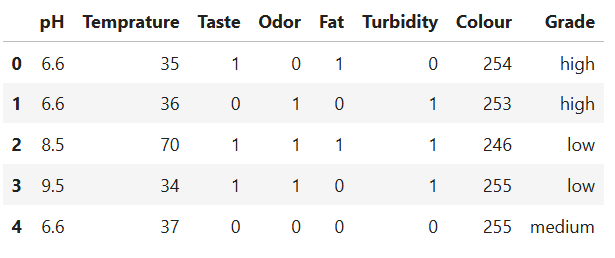
axs.set\_ylabel('accuracy rate')

axs.plot(depth\_vec, acc\_vec\_RF)

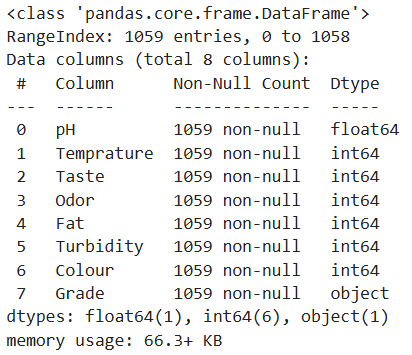
max(acc\_vec\_RF)

**CHAPTER 6**

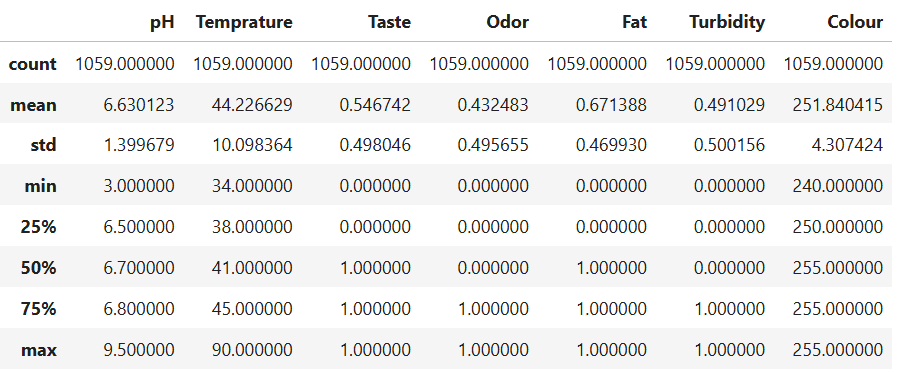
**RESULTS :**

****

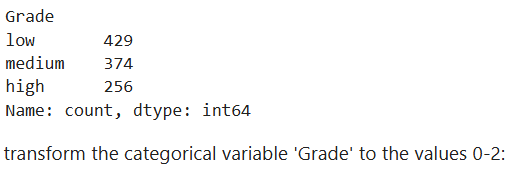
**Fig 6.1: Display of Dataset (first five).**

****

**Fig 6.2: Information about the dataset.**

****

**Fig 6.3: Display of the data distribution.**

****

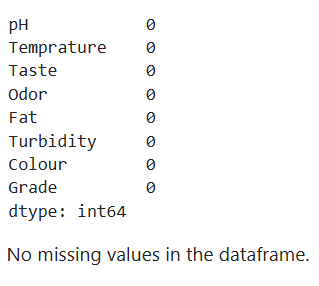
**Fig 6.4: Counts of Unique values in Grade column.**

This calculates the count of each unique value in the column Grade of the DataFrame df and returns it as a Pandas Series.Counts the occurrences of each unique value in the Grade column.Returns the counts in descending order by default.

****

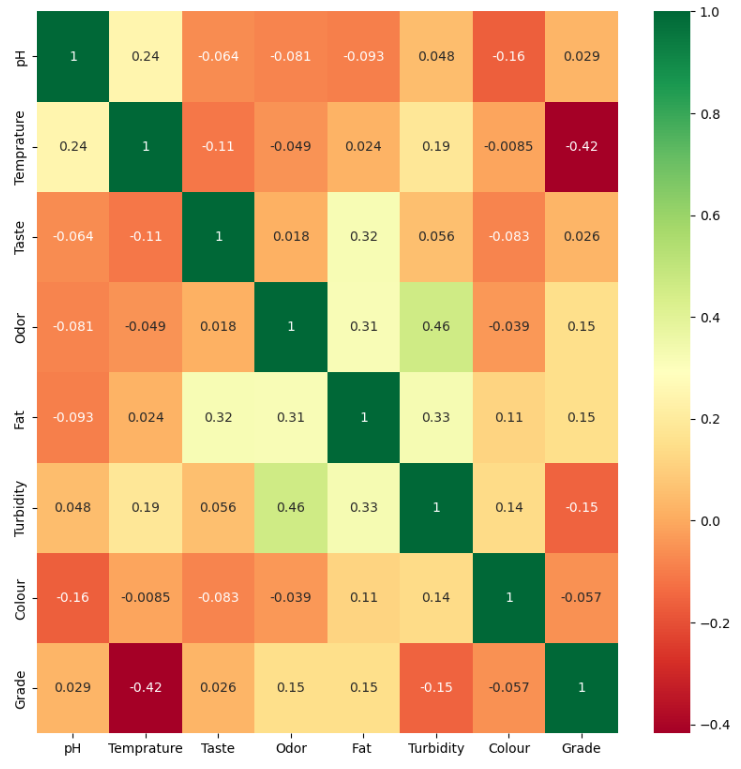
check for missing values:

**Fig 6.5: Updated Grade Values (high-2, medium-1, low-0).**

****

**Fig 6.6: Data of Null or Missing Values.**

It checks for missing values in the DataFrame df and returns the count of missing values for each column.

****

**Fig 6.7: Heatmap to visualize the correlations between numerical columns.**

**Breakdown of the Features:** pH: The data is concentrated around values between 6 and 8. A small number of outliers exist below 5 and above 8.

**Temperature:** The distribution is right-skewed, with most values between 40 and 50. Few samples have higher temperatures above 60.

**Taste:** Binary distribution (0 or 1). Indicates that the "Taste" feature is likely categorical.

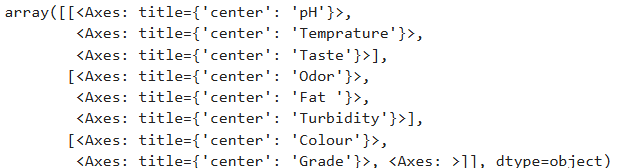
**Odor:** Binary distribution (0 or 1). Most samples fall under one category, showing an imbalance.

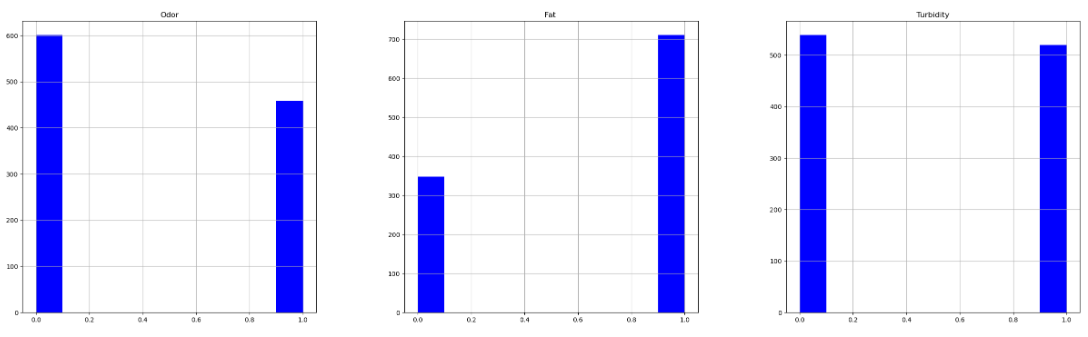
**Fat:** Binary distribution (0 or 1). Most samples belong to the 1 category, indicating a class imbalance.

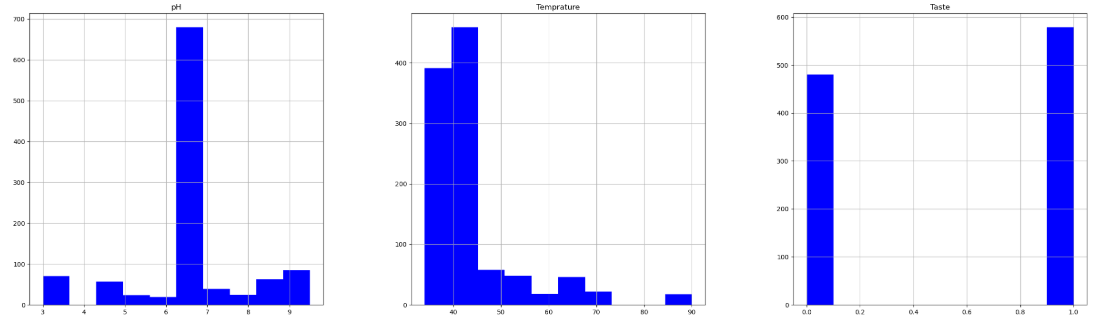
**Turbidity:** Binary distribution (0 or 1). Similar imbalance with most samples in one category.

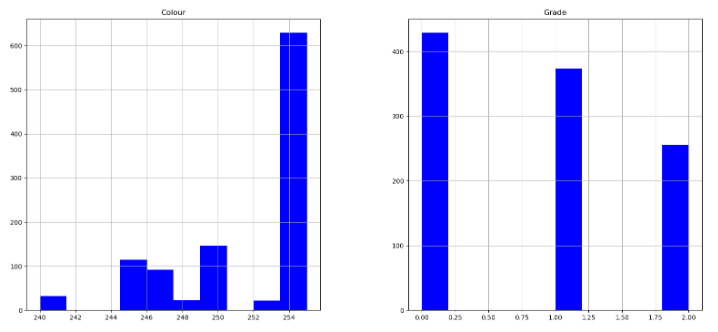
**Colour:** Appears to be continuous with values between 240 and 255. The data is heavily skewed toward the upper end, indicating limited variation.

**Grade:** Three distinct values (0, 1, 2), corresponding to low, medium, and high grades. The distribution shows more samples in lower grades compared to higher grades.



****

****

****

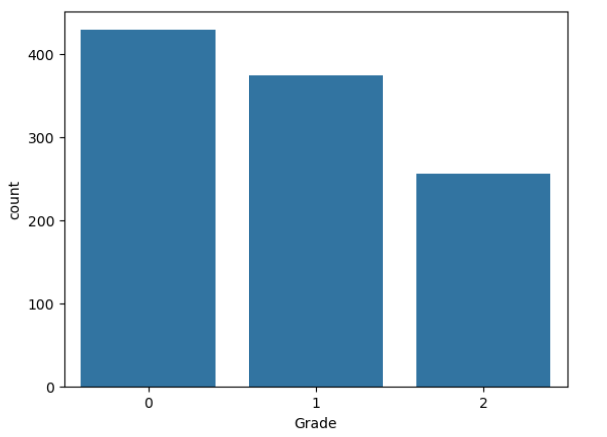
**Fig 6.8: Histograms for all the Numerical Columns.**

The provided bar chart shows the distribution of the Grade feature from the dataset. Here's what the chart represents:

**X-Axis (Grade):** The values 0, 1, and 2 correspond to the three levels of Grade: 0: Low grade 1: **Medium grade 2:** High grade

**Y-Axis (Count):** The height of the bars represents the number of samples belonging to each grade category.

**Observation:** Grade 0 (Low) has the highest count, indicating that most samples belong to this category. Grade 1 (Medium) is slightly less frequent. Grade 2 (High) is the least represented, showing an imbalance in the distribution of grades.

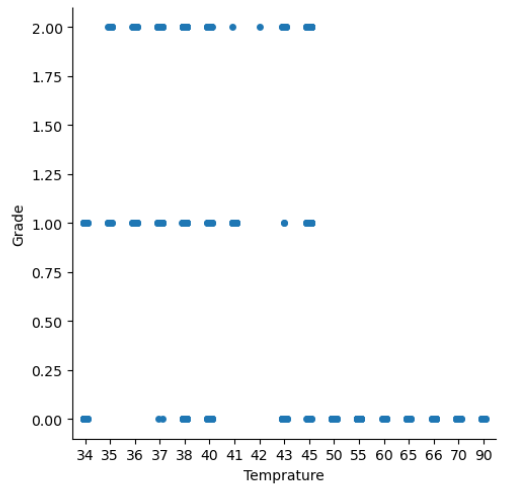
****

**Fig 6.9:** **Visualizing the Frequency Distribution of Grade.**

The chart represents the **count of milk quality grades** (e.g., Grade 0, 1, 2):

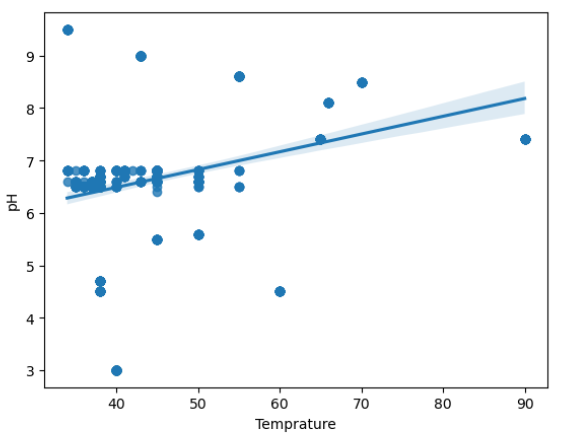
* **Grade 0**: Most samples fall under this category, indicating high-quality milk.
* **Grade 1**: Slightly fewer samples compared to Grade 0, showing medium quality.
* **Grade 2**: The least number of samples, representing lower-quality milk.

The output shows the distribution of quality levels, useful for identifying trends in milk quality and model performance.

****

**Fig 6.10: Strip plot to Visualize the relationship between Temperature and Grade.**

The scatter plot shows that as temperature increases, milk quality decreases. High-quality milk (Grade 0) is mostly found at lower temperatures, while medium (Grade 1) and low-quality milk (Grade 2) are associated with higher temperatures. This suggests that cooler temperatures help preserve milk quality.

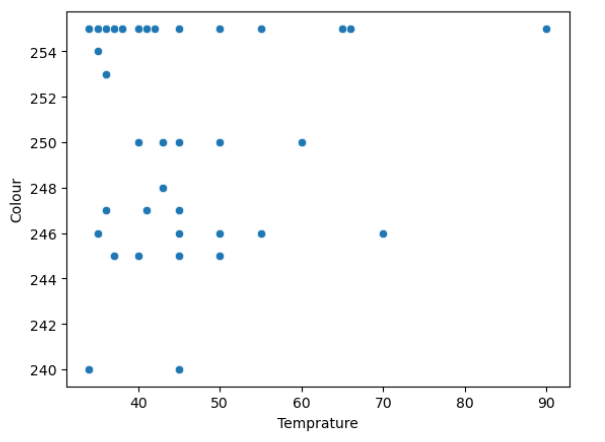
****

**Fig 6.11: Scatter Plot with a Regression Line to Visualize the relationship between Temperature and pH.**

This shows the relationship between temperature and pH values.

* **X-axis (Temperature)**: This axis represents the temperature values of the milk samples.
* **Y-axis (pH)**: This axis represents the pH levels of the milk samples.
* **Data Points**: Each point on the scatter plot corresponds to a milk sample's temperature and pH value.
* **Regression Line**: The line running through the points indicates a positive correlation between temperature and pH, meaning that as the temperature increases, the pH also tends to increase.
* **Confidence Interval**: The shaded area around the regression line represents the confidence interval, showing the range within which the true regression line is likely to fall.

This visual representation helps in understanding how temperature might influence the pH of milk, which is important for assessing milk quality.

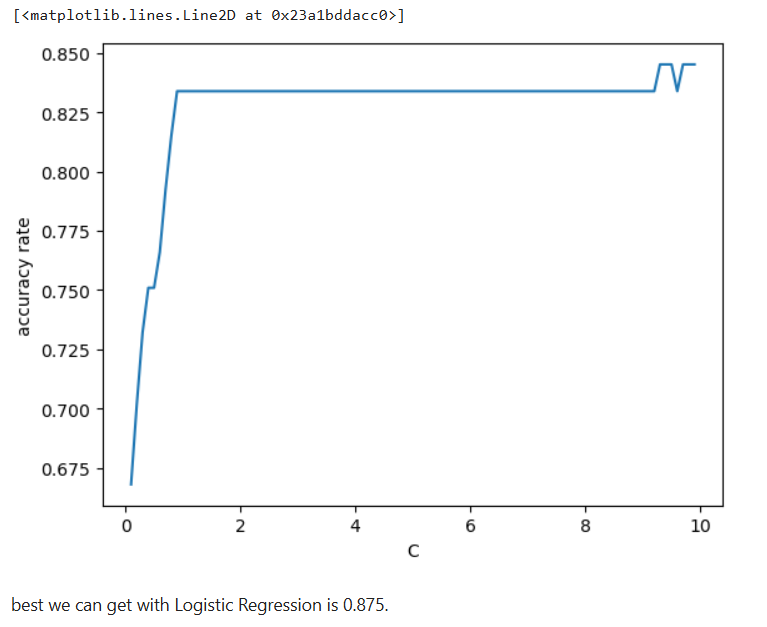
****

**Fig 6.12: Scatter Plot to Visualize relation between Temperature and Color.**

This graph is displays the relationship between temperature and color in the context of milk quality prediction.

* **X-axis (Temperature)**: This axis represents the temperature values of the milk samples, ranging from approximately 35 to 90.
* **Y-axis (Color)**: This axis represents the color values of the milk samples, ranging from 240 to 255.
* **Data Points**: Each point on the scatter plot represents a milk sample's temperature and color value.
* **Pattern**: The data points are scattered across the plot without a clear linear relationship, indicating that temperature and color might not have a straightforward correlation in this context.

This visualization helps in understanding how temperature might influence the color of milk, which could be a factor in predicting milk quality. It seems there isn't a strong relationship between these two attributes, suggesting other factors might be more significant in determining milk quality.

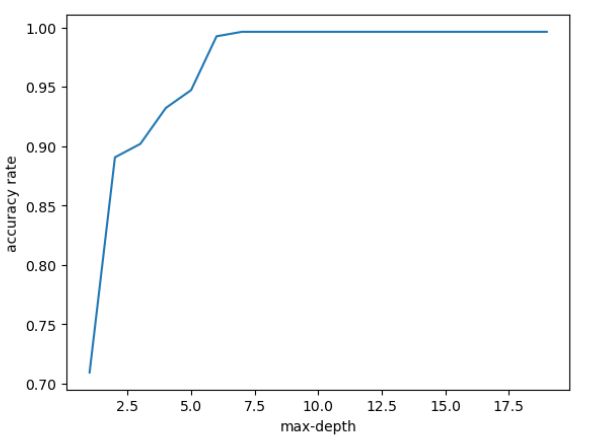
****

**Fig 6.13: Accuracy rate of Logistic Regression.**

The graph shows the relationship between the parameter C and the accuracy rate of your milk quality prediction model:

* **X-axis**: Different values of parameter C.
* **Y-axis**: Accuracy rate of the model.
* **Key Points:**
* **Accuracy Increases**: As C increases from 0 to around 2, the accuracy rate improves sharply.
* **Stabilization**: After C=2 , accuracy stabilizes around 0.825.
* **Fluctuations**: Slight fluctuations when C is around 10, but overall stability.
* **Conclusion:**

Increasing C improves accuracy up to C=2 , beyond which further increases have minimal effect.

****

**Fig 6.14: Accuracy rate of Random Forest.**

The graph shows how the accuracy of the milk quality prediction model changes with the maximum depth of a decision tree:

* **Accuracy Increases**: From max-depth 2.5 to 7.5, accuracy improves significantly.
* **Plateau**: Beyond max-depth 7.5, accuracy stabilizes at 100%.

**CHAPTER 7**

**CONCLUSION**

In conclusion, this project demonstrates the effectiveness of machine learning in predicting milk quality based on measurable parameters such as pH, temperature, taste, odor, fat, turbidity, and color. By integrating Logistic Regression and Random Forest models, the study provides a reliable framework for classifying milk quality into low, medium, and high grades. Among the models used, Random Forest proved to be more accurate and robust, making it an ideal choice for handling complex relationships in the dataset.

The methodology applied here is scalable and could be extended to include additional parameters or integrated with IoT-based systems for real-time data acquisition and prediction. This could revolutionize quality control in the dairy industry by providing faster, more accurate, and automated decision-making tools.

**CHAPTER 8**

**FUTURE SCOPE**

The future scope of this project lies in its potential to transform the dairy industry by introducing automation and precision into milk quality control processes. One key direction is the integration of this system with IoT devices and smart sensors, enabling real-time data collection and on-the-spot quality predictions during milk production and processing. This would not only enhance operational efficiency but also allow for proactive interventions to address quality issues.

The system could also be adapted to evaluate other parameters, such as bacterial counts or nutrient levels, offering a comprehensive tool for assessing milk quality. With the increasing focus on sustainable practices, this technology could play a role in minimizing waste by identifying and repurposing milk of suboptimal quality for other uses.

**REFERENCES**

[1]. HariniMaruthasalam, et al. Milk Quality Prediction Using Machine LearningTechniques. Repository. 2021.

[2]. Patel N, et al. Machine Learning Models for Milk Quality Prediction.Repository. 2020.

[3]. Randhir C, et al. Predicting Milk Quality Using Machine Learning.Repository. 2021.

**Website Referred:**

* <https://www.geeksforgeeks.org/machine-learning-projects/>
* [www.youtube.com](http://www.youtube.com)
* <https://github.com>/