MILK QUALITY PREDICTION

A Course Project Report submitted
in partial fulfillment of requirement for the award of degree

Bachelor of Technology

in

Computer Science & Engineering

by

S. Swathi	2103A51068
Hafsa	2103A51046
K. Tejaswini	2103A51561

Under the Guidance of
Mr. D. Ramesh
Assistant Professor, Department of CSE

DEPARTMENT OF COMPUTER SCIENCE & ARITIFICIAL INTELLIGENCE
SR UNIVERSITY, ANANTHASAGAR, WARANGAL

DEPARTMENT OF COMPUTER SCIENCE & ARTIFICIAL INTELLIGENCE

CERTIFICATE

This is to certify that the Project entitled "**Milk Quality Prediction**"is a bona fide work carried out by the **S. Swathi, Hafsa, K. Tejaswini** bearing Roll No(s) **2103A51068, 2103A51046, 2103A51561** as a Course Project for the partial fulfillment to award the degree of **Bachelor of Technology** in **Computer Science Engineering** during the academic year 2022-2023 under our guidance and supervision.

Mr. D. Ramesh
Asst. Professor
SR University,
Ananthasagar, Warangal.

Head of the Department

ACKNOWLEDGEMENT

First and foremost, we express our sincere thanks for the guidance and encouragement rendered, D. Ramesh in the Department of Computer Science & Artificial Intelligence, SR University, Ananthasagar, Hanamkonda District. We extend our gratitude for his advice and guidance during the progress of this course project.

Secondly, we express our sincere thanks to **Dr. M. Shashikala, Associate Professor & Head**, Department of CS & AI, SR University who stood as silent inspiration behind this course project. Our heartfelt thanks for her endorsement and valuable suggestions.

We wish to express our profound thanks to **Dr. C. V. Guru Rao, Registrar & Dean, School of CS & AI** for providing necessary facilities to make this course project a success.

We thank all the members of teaching and non-teaching staff members, and also who have assisted us directly or indirectly for successful completion of this course project.

Finally, we would like to express our sincere gratitude to our parents who are constantly encouraging us through-out our lives and for completion of this course project.

2103A51068 - S. Swathi
2103A51046 - Hafsa
2103A51561 - K. Tejaswini

ABSTRACT

Milk quality prediction is an important area of research that aims to improve the overall quality of milk by identifying potential factors that may affect its quality. With the increasing demand og high quantity of milk, it has become necessary to develop accurate and reliable methods for predicting milk quality. Various techniques have been used in milk quality prediction, including sensory evaluation, chemical analysis, and microbial analysis. These methods are time consuming, expensive, and may not be suitable for large scale dairy farms.

Recently, machine learning based approaches have emerged as a promising solution to predict milk quality. These methods use data from various sources. Machine learning algorithms such as decision tree, neural networks, and support vector machines have been used to develop predictive models for milk quality. Machine learning based approaches offer a promising solution for predicting milk quality, which can help dairy farmers take appropriate actions to improve milk quality and meet the increasing demand for high quality milk.

Table of Contents

CHAPTER 1

INTRODUCTION

The quality of the milk has been a concern for a long time. Adulteration of milk with water, urea and other substances is a common practice in some areas of the country. This practice reduces the nutritional value of milk and also poses health risks to consumers.

The government of India has introduced several measures in order to address the issue. Including the Food Safety and Standards Authority of India(FSSAI), which sets standards for milk and milk products and various programs to promote the production of safe and high quality milk. Moreover, various initiatives have been taken by private dairy companies and cooperatives to ensure the quality of milk, including the introduction of modern milking and storage equipment, setting up milk collection centers and implementing quality testing procedures. And even consumers also choose to buy branded or packaged milk that meets the FSSAI standards.

Predicting milk quality can be essential for ensuring that the milk produced is safe for consumption and meets necessary standards. It can also help in identifying potential issues in the production process and taking appropriate correct measures.

1.1 Problem Statement

One of the major challenges of the dairy industry is ensuring the quality of milk. Milk quality prediction is a critical task that aims to detect and prevent adulteration of milk, which can cause several health issues for the consumers. The problem of milk quality involves developing accurate and efficient models that can predict the quality of milk based on various parameters such as fat content, temperature, pH of milk etc.

Milk quality predictions can be particularly important in the dairy industry, where large amounts of milk are produced and processed daily. Milk with poor quality can have a negative impact on consumer health and can lead to economic losses for the dairy industry. Not only the adulteration process but the quality of milk can be also affected by various other factors such as the health and diet of the cows, the handling and storage conditions and the processing techniques used.

1.2 Existing System

Existing methods for milk quality prediction includes traditional laboratory-based techniques, which are time consuming and expensive. However, this methods may not be possible for small-scale

dairy farmers, who lack the resources to invest in such technologies. Therefore, there is a need for innovative technique which can give accurate results and can be used by all types of dairy farmers. To develop a milk quality prediction model, machine learning algorithms can be employed. The model can be trained on a dataset containing milk quality measurements and their corresponding parameters to evaluate its accuracy.

1.3 Proposed System

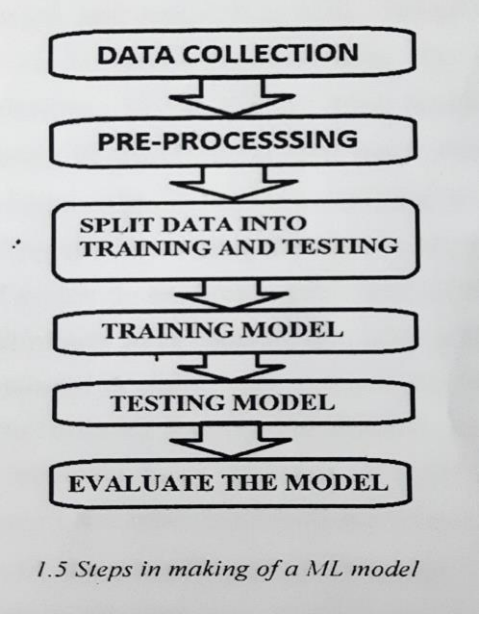
With the assist of dataset obtained we create different algorithms, specifically Logistic Regression, KNN, Decision Tree etc. and examine the outcomes of accuracy and find which models performs better and is reliable.

1.4 Objectives

- Compare the accuracy in 5 specific classification-based system learning algorithms.
- To establish machine learning algorithms are reliable for automatic results.
- These algorithms can be used to make predictions on new data, allowing for real-time monitoring and early detection of new issues.

1.5 Architecture

The Supervised Learning approach as a qualitative data with KNN classification, SVM, Decision tree, Bayes theorem, Logistic Regression and its target to predict safe and quality of the milk.



1.5 Steps in making of a ML model

CHAPTER 2

LITERATURE SURVEY

In previous research papers, we have observed that different machine learning algorithms have been used. Few papers are based on deep learning also. The field of Artificial Intelligence has been the suitable area to carry out all types of predictions on the dataset by extracting and data preprocessing. Logistic Regression, Support Vector Machine, Naïve Bayes Classification, Linear regression and ridge regression etc. are the various machine learning algorithms the have been used.

We have observed that the algorithms work together by generating the pattern among the available dataset and proceeding with prediction. Mid Infrared Spectroscopy combined with few machine learning algorithms. Deep learning is something that works by generating biases and weights in the layers, rule based takes the bulk values and signifies a rule in it. SVM are used with algorithms especially which follows a close correlation among the variables taken into consideration.

Artificial Neural Network inspired by the structure and function of the human brain. PLS regression stands for Partial Square regression, which is a statistical technique used for modelling the relationship between the two sets of variables. In PLS regression, both the predictor variables and the response variables are transformed into new sets of variables called latent variables, which are linear combination of the original variables. PLS regression is useful for predicting a response variable from a large number of predictor variables, even when these variables are highly correlated. It is commonly used in fields such as chemistry, biology, and engineering, where there are many variables to consider in modelling complex systems. It is also used in data analysis and machine learning to identify important variables and reduce dimensionality of the data.

SI NO	DATE OF PUBLICATION	AUTHORS	NAME	METHODOLOGY	ACCURACY
1	2018/8/1	Dixon Vimalajeewa	Learning in the compressed data domain: Application to milk quality prediction	MIRS	92.3 91.1
2	2015/9/1	G Visentin	Prediction of dairy milk technological traits from mid-infrared spectroscopy analysis in dairy cows	Combination with MIRS and PLS regression	77 86
3	2009/5/1	RRB Singh	Prediction of sensory quality of UHT milk	ANN, Chemical Kinetics	92 94
4	2021/7/1	M Frizzarin	Predicting cow milk quality traits from routinely available milk spectra using statistical and machine learning methods	PLSR NN	91.57 67.86
5	2015/9/1	Donagh Berry	Prediction of Bovine milk technological traits from mid-infrared spectroscopy analysis in dairy cows	Combination of MLSR and PLS Regression	70.76
6	2009/1/1	Valentina Bonafatti	Prediction of protein composition of individual cow milk using mid infrared spectroscopy	RP HPLC	75
7	2017/8/30	Clement Grelet	Prediction of energy status of dairy cows using MIR milk spectra	PLS, SVM	84 81
8	2021/2/1	JA Fernandez	Large scale Phenotyping in dairy sector using milk MIR spectra	PLS	86

CHAPTER 3

DATA PRE-PROCESSING

Data pre-processing is an important step in preparing data for analysis and modeling. When it comes to milk quality prediction, there are several steps involved in data pre-processing. They are data collection, data cleaning, data transformation, feature selection, data splitting, encoding categorical variables, handling outliers. By carefully cleaning, transforming, and selecting features from the data, we can build a more accurate and effective model for predicting milk quality.

3.1 Dataset Description

The dataset for milk quality prediction includes various parameters related to milk quality. These parameters can be physical and chemical properties of milk. This dataset contains 1059 rows of data and 7 columns of different parameters related to milk. There is another column grade which is the overall result of the data and it is also the target variable of the dataset. The seven parameters of the milk include pH, Temperature, Taste, Odor, Fat, Turbidity, Color. Each sample of data includes all these seven parameters.

	A	B	C	D	E	F	G	H
1	pH	Tempratur	Taste	Odor	Fat	Turbidity	Colour	Grade
2	6.6	35	1	0	1	0	254	1
3	6.6	36	0	1	0	1	253	1
4	8.5	70	1	1	1	1	246	3
5	9.5	34	1	1	0	1	255	3
6	6.6	37	0	0	0	0	255	2
7	6.6	37	1	1	1	1	255	1
8	5.5	45	1	0	1	1	250	3
9	4.5	60	0	1	1	1	250	3
10	8.1	66	1	0	1	1	255	3
11	6.7	45	1	1	0	0	247	2
12	6.7	45	1	1	1	0	245	2
13	5.6	50	0	1	1	1	255	3
14	8.6	55	0	1	1	1	255	3
15	7.4	90	1	0	1	1	255	3
16	6.8	45	0	1	1	1	255	1
17	6.5	38	1	0	0	0	255	2
18	4.7	38	1	0	1	0	255	3
19	3	40	1	1	1	1	255	3
20	9	43	1	0	1	1	250	3
21	6.8	40	1	0	1	0	245	2
22	6.6	45	0	1	1	1	250	1
23	6.5	36	0	0	1	0	255	2
24	4.5	38	0	1	1	1	255	3
25	6.6	45	1	1	1	1	245	1
26	6.8	35	1	0	1	0	246	2

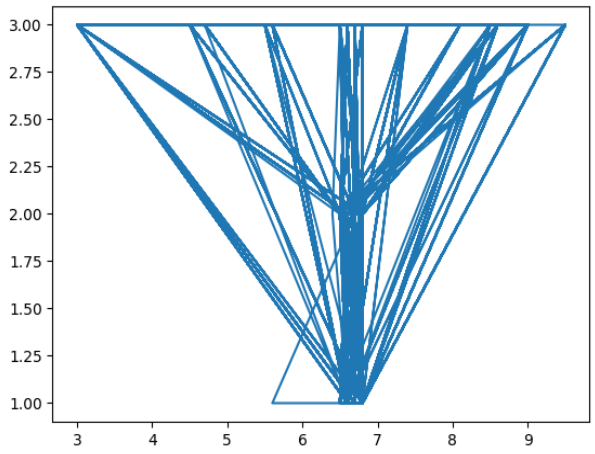
The target variable for the prediction model is a classification indicating whether the milk is of high or medium or low quality. The dataset is provided in a CSV format and can be used to train and evaluate machine learning models for milk quality prediction. The fat attribute gives the fat content present in the sample of the milk, turbidity is the cloudiness or opacity of milk caused by the presence of suspended particles in the sample of the milk. By using these seven factors high, low, medium quality classification is carried out.

3.2 Pre-Processing through standard scaler

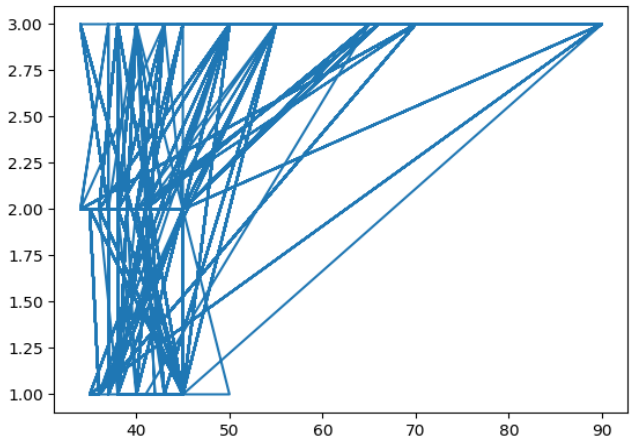
Here in our dataset, we used target encoding, which involves encoding categorical variables based on target variable. In our dataset, at the grade attribute instead of high, medium, low quality we used 1, 2, 3 numbers respectively. This method can improve model performance.

3.3 Data Visualization

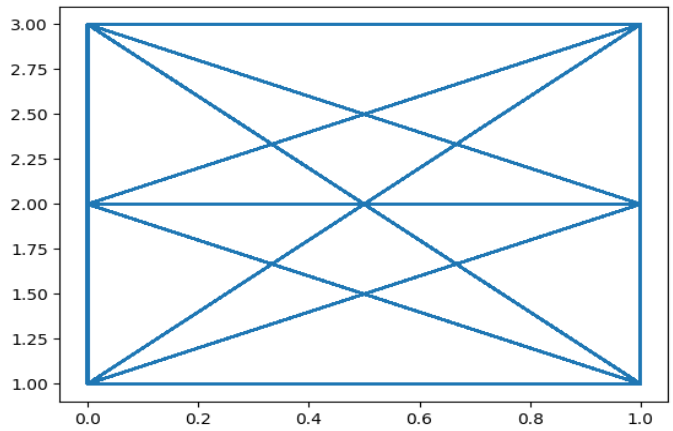
The following are the plotting of each attribute against the target.



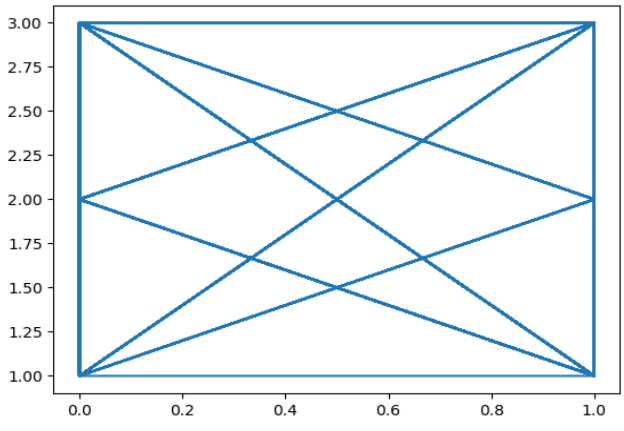
pH Vs Grade



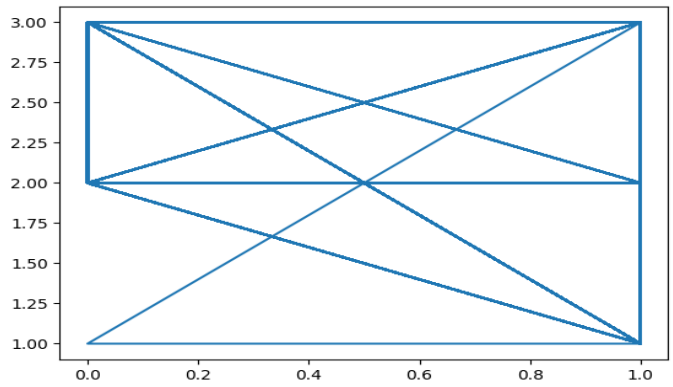
Temperature Vs Grade



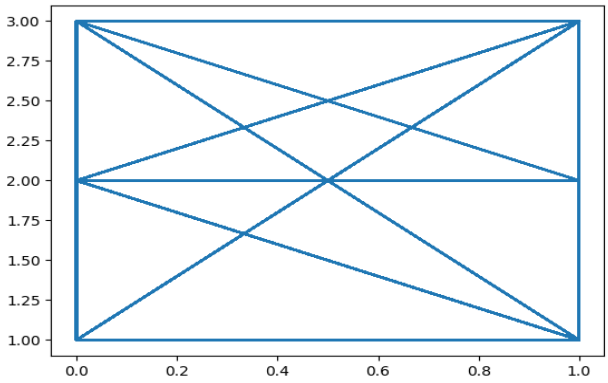
Taste Vs Grade



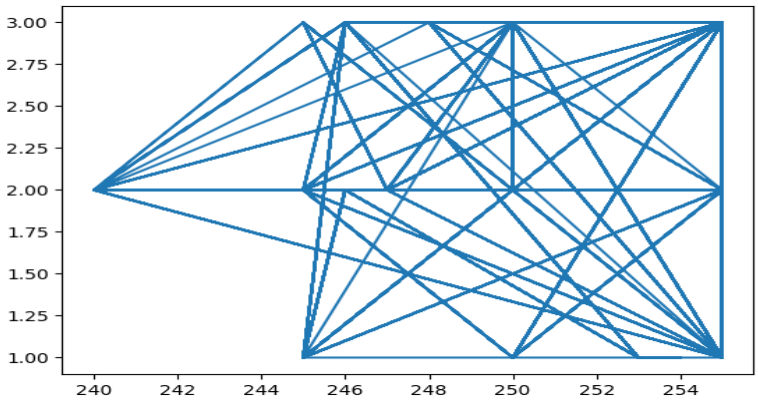
Odor Vs Grade



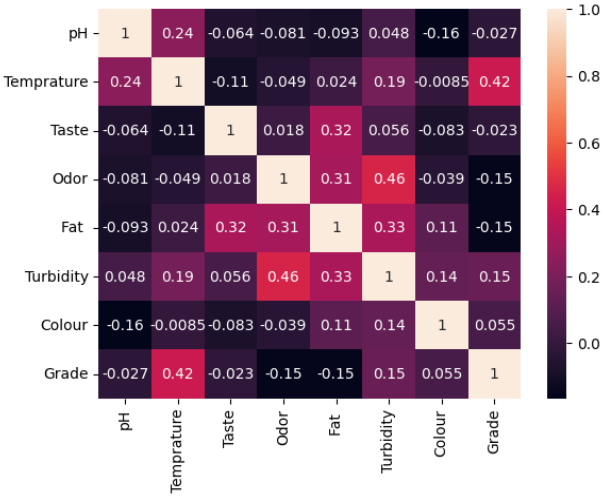
Fat Vs Grade



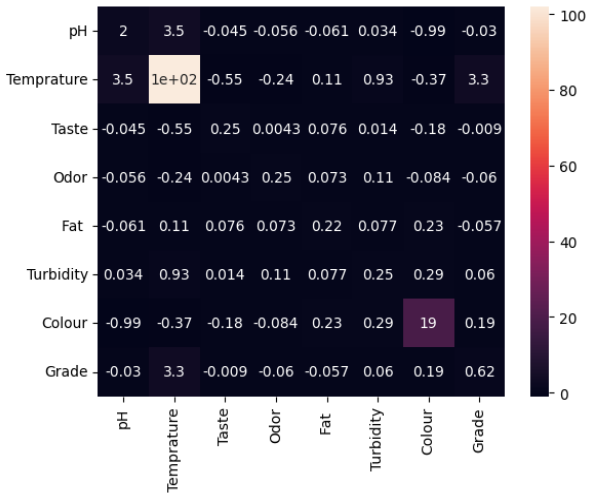
Turbidity Vs Grade



Color Vs Grade



Correlation Matrix



Covariance Matrix

- ✓ A Correlation matrix is a table that displays the correlation coefficients between multiple variables.
- ✓ A correlation coefficient 1 indicates a perfect positive correlation, coefficient 0 indicates no correlation and coefficient -1 indicates a perfect negative correlation.
- ✓ By analyzing the correlation matrix, researchers and analysts can identify which factors have the strongest correlation with the milk quality and use that information to develop models or interventions to improve milk quality.
- ✓ A Covariance matrix is a square matrix that contains the covariances between pairs of variables in a dataset.
- ✓ The diagonal elements of a covariance matrix represent the variance of each variable, while the off-diagonal elements represent the covariance between pairs of variables.
- ✓ A positive covariance between two variables indicates that they tend to increase or decrease together, while a negative covariance indicates that one tends to while the other decreases.
- ✓ By analyzing the covariance matrix, researchers and analysts can identify which factors have the strongest covariance with the milk quality and use that information to develop models or interventions to improve milk quality.

CHAPTER 4

METHODOLOGY

4.1 Logistic Regression

Logistic Regression uses the concept of predictive modeling as regression; therefore it is also called logistic regression, but it is used to classify samples therefore it falls under classification algorithm. In other words logistic regression estimates the probability that an event occurs based on the values of one or more independent variables.

Logistic Regression is a commonly used algorithm in various fields, including finance, marketing, healthcare and social sciences. It is a simple and efficient algorithm that can provide accurate predictions, but it is sensitive to outliers and may require careful preprocessing.

Logistic Regression uses a sigmoid or logistic function which will squash the best fit straight line that will map any values including the exceeding values from 0 to 1 range. So, it forms as an "S" shaped curve. Sigmoid function removes the effect of outlier and makes the output between 0 to 1.

```
from sklearn.linear_model import LogisticRegression
lr=LogisticRegression()
mm=lr.fit(x_train,y_train)
```

4.2 Decision Tree

Decision trees are a nonparametric supervised learning method used for classification and regression. The deeper the tree, the more complex the decision rules and the fitter the model. Decision tree uses the tree representation to solve the problem. In which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree. The primary challenge in the decision tree implementation is to identify the attributes. There are two popular attribute selection measures they are Entropy and Gini index.

Entropy is the measure of uncertainty of a random variable, it characterizes the impurity of an arbitrary collection of examples. The higher the entropy more the information content.

$$Entropy = \sum_{i=1}^{C} -p_i * \log_2(p_i)$$

Information Gain is a measure of the change in entropy.

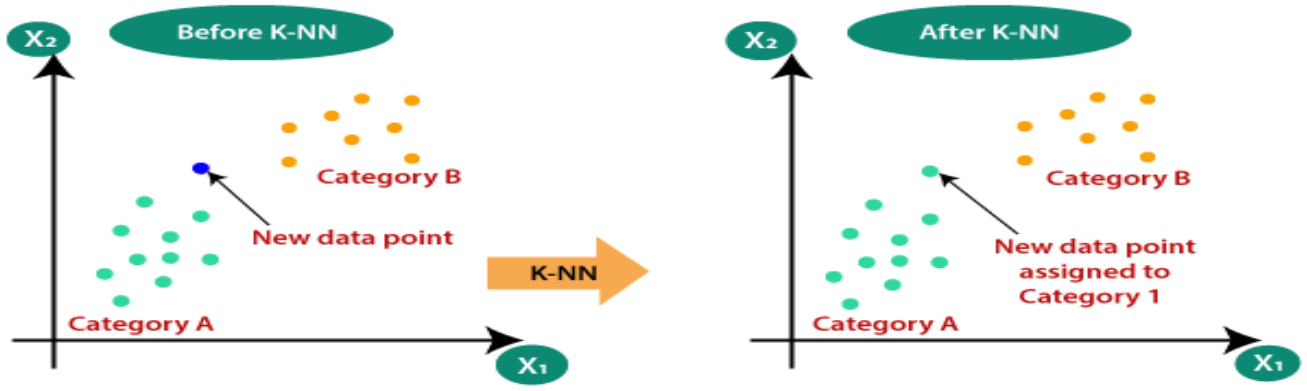
$$\text{Information Gain}(T,X) = \text{Entropy}(T) - \text{Entropy}(T, X)$$

```
from sklearn.tree import DecisionTreeClassifier
classifier=DecisionTreeClassifier(criterion='entropy',random_state=0)
mm=classifier.fit(x_train,y_train)
```

4.3 K-Nearest Neighbor

K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into well suite category by using K-NN algorithm.

1. Select the value of K in the K-NN algorithm.
2. Calculate the Euclidean distance of K number of neighbors.
3. Take the K nearest neighbors as per the calculated Euclidean distance.
4. Among this K neighbors, count the number of the data points in each category.
5. Assign the new data points to that category for which the number of the neighbor is maximum.




```
from sklearn.neighbors import KNeighborsClassifier
classifier=KNeighborsClassifier(n_neighbors=5,metric='minkowski',p=2)
classifier.fit(x_train,y_train)
```

4.4 Gaussian Naive Bayes

Gaussian Naïve Bayes is a probabilistic algorithm used for classification tasks. It is based on bayes theorem and assumes that the probability of a feature belonging to a certain class is independent of the values of other features.

The algorithm assumes that the probability distribution of each feature is Gaussian and estimates the mean and standard deviation of each feature for each class in the training data. During prediction, the algorithm calculates the probability of each class for a given set of features using Bayes' theorem and selects the class with the highest probability.

Gaussian Naïve Bayes is particularly useful for high dimensional datasets, as it can efficiently handle a large number of features with relatively small amounts of training data. However, its assumption of independence between features may not hold true in some cases, leading to suboptimal performance.

```
from sklearn.naive_bayes import GaussianNB
gnb=GaussianNB()
gnb.fit(x_train,y_train)
```

4.5 Support Vector Machine

In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is the number of features you have). We perform classification by finding the hyperplane that differentiates the two classes very well. The distance between the hyperplane and the nearest data point from either set is known as the margin. The goal is to choose a hyperplane with the greatest possible margin. There will never be any data point inside the margin.

One of the advantages of SVM is its ability to handle high dimensional data with a small sample size, as well as its robustness to noise and outliers. However, SVM can be computational expensive, especially for large datasets. In addition, SVM requires careful selection of hyperparameters, such as the kernel function and regularization parameter, to achieve optimal performance.

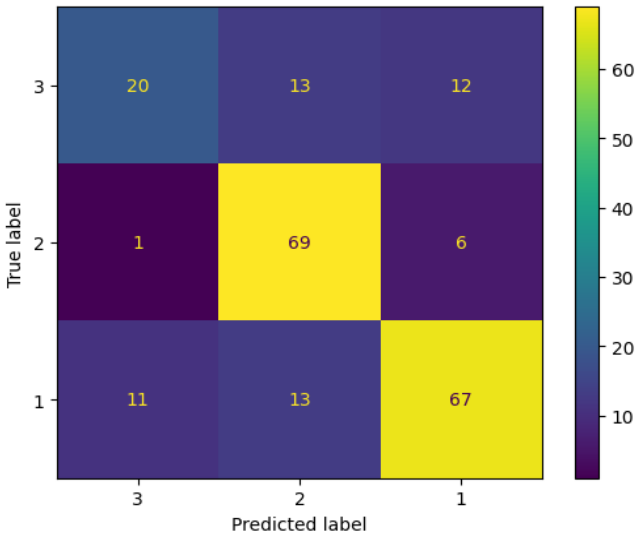
```
from sklearn.svm import SVC
svm_model=SVC(kernel='linear')
svm_model.fit(x_train,y_train)
```

CHAPTER 5

RESULTS

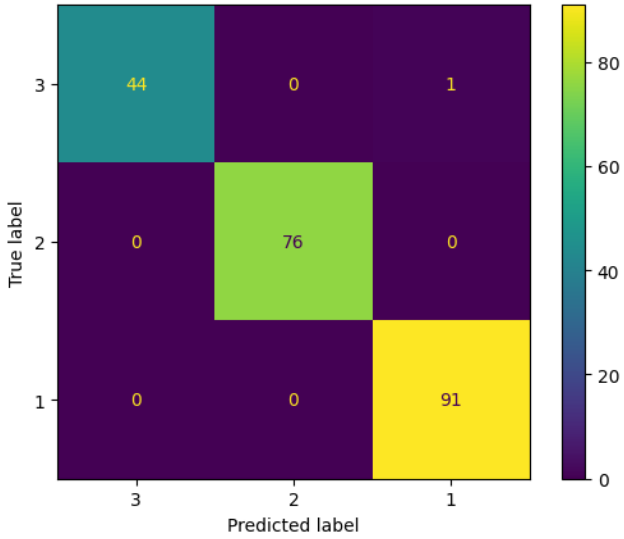
Logistic Regression

Accuracy: 0.7358490566037735



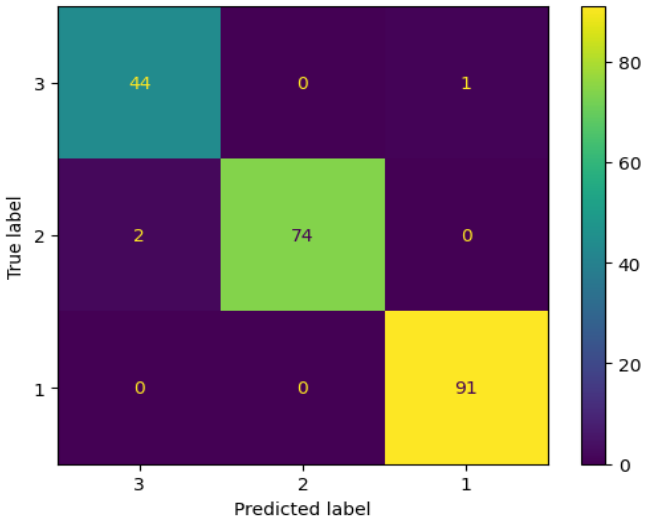
Decision Tree

Accuracy: 0.9952830188679245



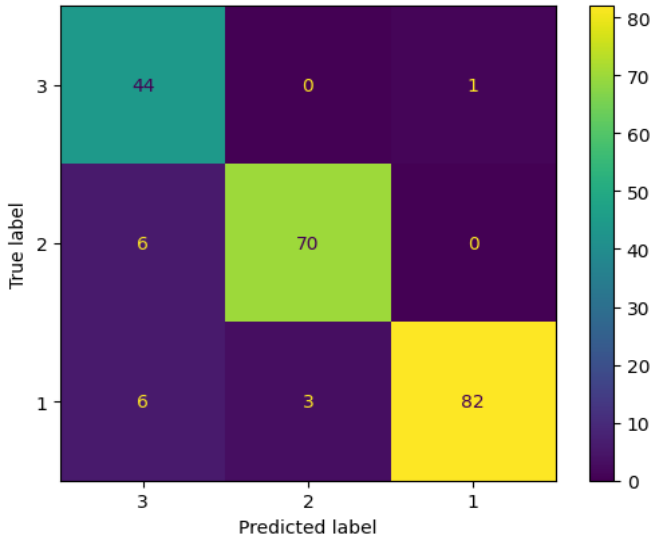
K-Nearest Neighbor

Accuracy: 0.9858490566037735



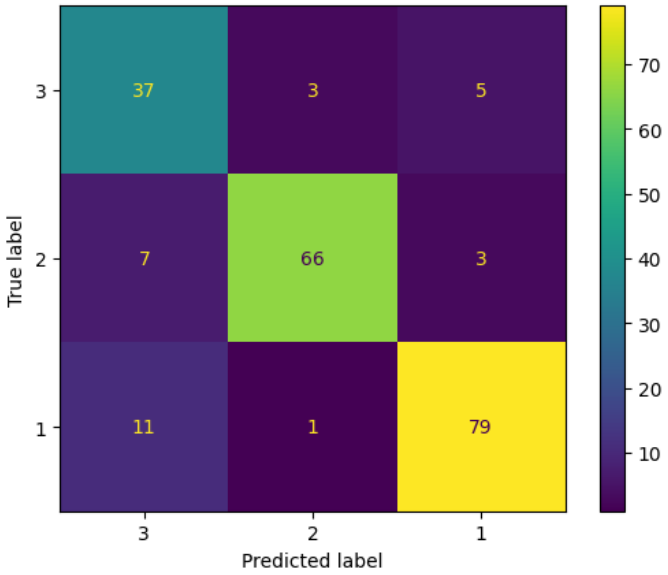
Gaussian Naïve Bayes

Accuracy: 0.9245283018867925



Support Vector Machine

Accuracy: 0.8584905660377359



CHAPTER 6

CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

Finally, after performing all the steps needed to get the results from preparation to preprocessing to performing the models (Logistic Regression, Decision tree, K-Nearest Neighbor, Gaussian Naïve Bayes, Support Vector Machine)**.** If we observe, the percentage of accuracy of Logistic Regression, Decision tree, KNN, Gaussian Naïve Bayes, SVM are 73.58, 99.52, 98.58, 92.45, 85.84 respectively. Out of which Decision tree model with 99.52830188679245 percent accuracy performs relatively better than all other models and secondly KNN model performs better with accuracy percentage of 98.58490566037735.

6.2 Future Scope

The future scope of milk quality prediction is promising, as there is growing demand for high quality milk in the global market. The use of machine learning based approaches and advanced technologies is expected to play a crucial role in improving the accuracy and efficiency of milk quality prediction.

The future scope of milk quality prediction is bright, and the development of advanced technologies and innovative approaches is expected to contribute significantly to the improvement of milk quality and the sustainability of the dairy industry.

REFERENCES

[1] http://repository.wit.ie/3326/1/InfomationScience_postprint.pdf

[2] https://www.sciencedirect.com/science/article/pii/S0022030215004932

[3] https://www.sciencedirect.com/science/article/abs/pii/S0260877408005323

[4] https://www.sciencedirect.com/science/article/pii/S0022030221005099

[5] https://www.sciencedirect.com/science/article/pii/S0022030215004932

[6] https://www.tandfonline.com/doi/abs/10.4081/ijas.2009.s2.399

[7] https://orbi.uliege.be/handle/2268/224000

[8] https://www.sciencedirect.com/science/article/abs/pii/S1046202320301158