TITLE:- PREDICT THE MILK QUALITY USING MACHINE LEARNING ALGORITHMS

Date: -15-02-2023
Title: predict milk quality using ML algorithm
Mail id:-chetanahirrao2022@gmail.com
Mob:-8408004897
GITHUB:- https://github.com/chetan456789
LINKEDIN:- https://www.linkedin.com/in/chetan-ahirrao-602305b5/

Abstract:

In today's highly competitive production industry, customer satisfaction is a major concern for service providers. Customer churn occurs when customers switch to a competitor's service, resulting in lost revenue for the provider. In this project, for predicting milk quality and improving the quality and safety of dairy products. we aim to develop a machine learning model that can predict customer churn, thereby enabling the provider to take proactive measures to retain customers.

We will use a dataset that contains milk ingredients information such as PH, turbidity, colour, odour. We will preprocess the data by cleaning it and performing feature engineering to extract meaningful information from the raw data. We will then split the data into training and testing sets to train and evaluate our machine learning model.

We will explore different classification algorithms such as Naïve bayes, K-Nearest neighbour, decision trees, and support vector machines to determine which algorithm provides the best accuracy in predicting customer churn.

Finally, we will evaluate the performance of our model using metrics such as accuracy, precision, recall, and F1-score. We can also perform a cost-benefit analysis to estimate the financial impact of our model on the provider's business.

Overall, this project aims to demonstrate the effectiveness of machine learning in predicting customer churn in the service sector industry and to provide insights to the provider for customer retention strategies.

1.introduction

predicting milk quality is essential for ensuring consumer safety, meeting industry standards, reducing waste, and improving profitability. Machine learning algorithms can be a valuable tool for predicting milk quality and improving the quality and safety of dairy products.

Ensuring consumer safety: Milk is a perishable product that can spoil quickly, and milk of poor quality can contain harmful bacteria or other contaminants that can cause illness. By predicting milk quality, dairy farmers and milk processors can ensure that only high-quality milk is sold to consumers, reducing the risk of foodborne illness.

Meeting industry standards: Many countries have strict regulations and standards for the quality of milk, including minimum levels of fat, protein, and other nutrients. By predicting milk quality, dairy farmers and milk processors can ensure that their products meet these standards and avoid penalties or fines for non-compliance.

Reducing waste: Milk of poor quality can be discarded, resulting in significant waste and financial losses for dairy farmers and milk processors. By predicting milk quality, dairy farmers and milk processors can identify milk that is likely to be of poor quality and take steps to prevent it from entering the food supply, reducing waste and costs.

Improving profitability: High-quality milk is more valuable than low-quality milk, and dairy farmers and milk processors can command higher prices for their products if they can demonstrate their quality. By predicting milk quality, dairy farmers and milk processors can improve the overall quality of their milk and increase their profitability.

1.1 About the data

Predicting milk quality using machine learning algorithms involves building a model that can classify milk samples into different quality categories based on their characteristics. Some of the key steps involved in building a milk quality prediction model are:

Data collection: Collect data on different milk samples, including their quality characteristics such as fat content, turbidity, pH, Taste, Odor, Color, temperature and other chemical properties.

Data pre-processing: Clean and pre-process the data to remove any missing or inconsistent values, normalize the data, and prepare it for analysis.

Feature selection: Identify the most important features that can help predict the quality of milk. This can be done using statistical analysis or machine learning techniques.

Model training: Select an appropriate machine learning algorithm, such as logistic regression, decision tree, random SVC, and train the model on the data. The model is trained to predict the quality category of milk samples based on their characteristics.

Model evaluation: Evaluate the performance of the model on a separate test dataset to determine its accuracy, precision, recall, and F1 score.

Model deployment: Once the model has been trained and evaluated, it can be deployed in a production environment to predict the quality of new milk samples.

Overall, predicting milk quality using machine learning algorithms is a complex process that requires expertise in data analysis, machine learning, and domain knowledge of milk quality characteristics. It can be a valuable tool for dairy farmers and milk processing companies to improve the quality of their products and reduce costs.

DATA EXPLORATION ANALYSIS

Here we have done some exploratory data analysis on given dataset we have Plotted box plot of grade vs turbidity to check any outliers in given dataset.

```
ax = sns.boxplot(x='Grade', y='Turbidity', data=df)
plt.xticks(rotation=45)
plt.title('Grade v/s Turbidity')
```

```
Out[14]: Text(0.5, 1.0, 'Grade v/s Turbidity')
```

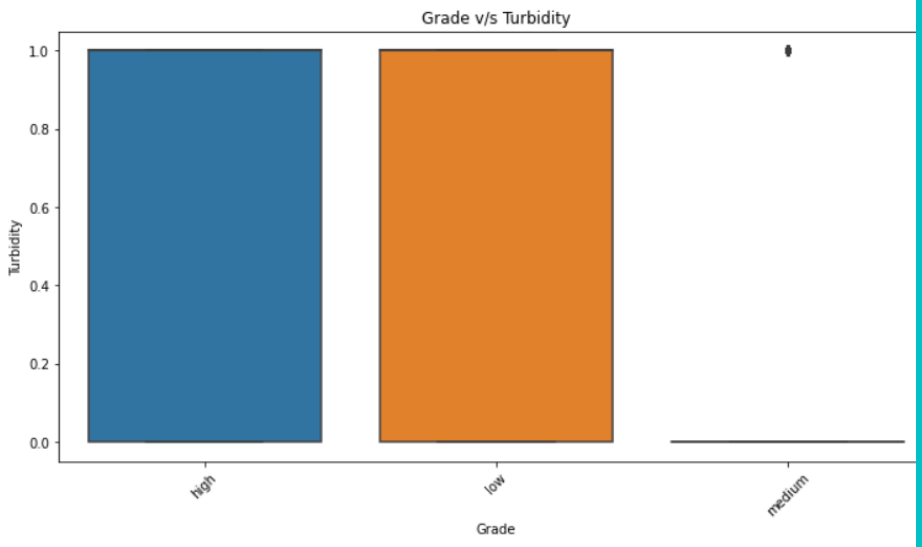

Figure1: grade vs turbidity

Plot of grade vs odor.

```
In [16]: plt.figure(figsize = (12, 6))
         ax = sns.boxplot(x='Grade', y='Odor', data=df)
         plt.setp(ax.artists, alpha=.5, linewidth=2, edgecolor="k")
         plt.xticks(rotation=45)
         plt.title('Grade v/s Odor')

Out[16]: Text(0.5, 1.0, 'Grade v/s Odor')
```

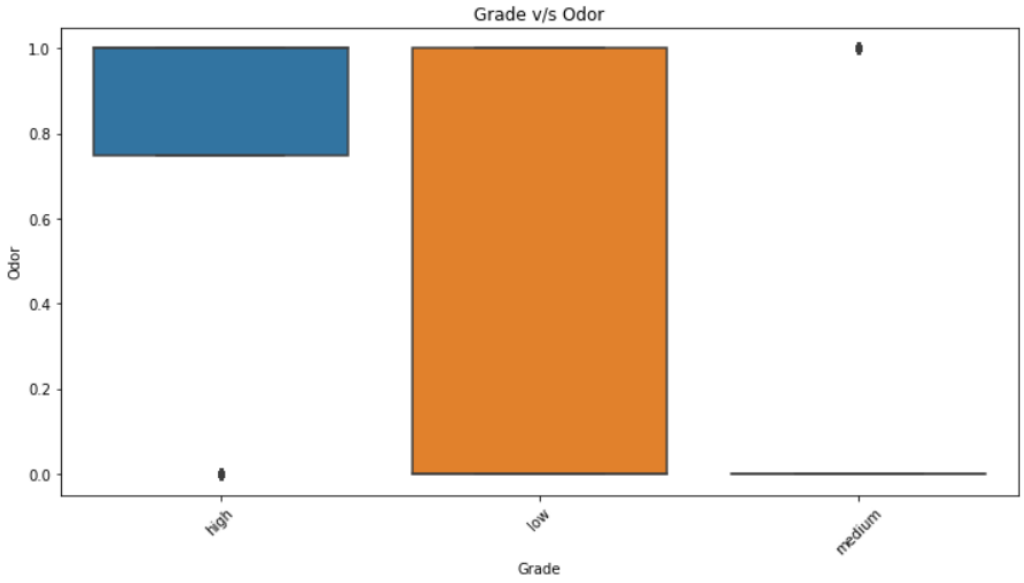

Figure 2: grade vs odor

As we can analyses max value is 1 and min value is -1 so which parameter is contributing and affecting the grade of the milk higher the value of the fat contents leads to lower the grade value of the milk colors represents the intensity of the data. if the ph values increases then basic property of the milk and if the ph value is decreases the more acidic property of the milk will increases.

```
In [22]: corr=df.corr()
         plt.figure(figsize=(12,8))
         sns.heatmap(corr,vmax=1.0, vmin=-1.0, cmap='viridis', annot=True)
Out[22]: <AxesSubplot:>
```

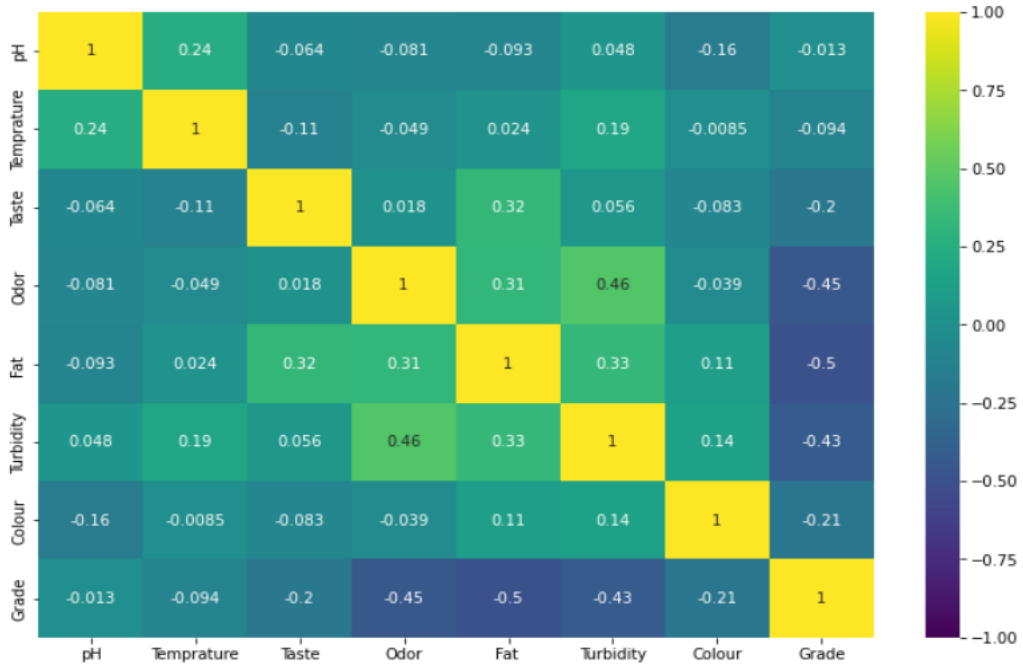

Figure 3: Heatmap of dataset

3 Experiments:

if we have categorical variable for ex- large, small then we cannot directly provide it to machine it won't be able to understand so ml algorithm involves lots of mathematical calculation so in this case we will use label encoder because we have order features in target variable Grade like high low and medium. In our given dataset there is no any null values so we don't need to drop that missing values.

At this stage, completes the data cleaning and manipulation process.

3.1 Evaluation metrics

After doing exploratory data analysis the next step is evaluation metric where we build the model and apply that model to our given dataset that means how well a model is able to make predictions or classifications based on a set of input data.

There are wide variety of evaluation metrics that can be used depending upon the specific task we have use some of them like below.

1. *Accuracy: - measures the proportion of correctly classified instance. accuracy summaries the whole model. it is the ratio of correctly classifies prediction to the entire predictions.*

   Accuracy=corrected predictions /All predictions

2. *Precision: - measures the proportion of true positive (correctly identified instances) among all the instances predicted as positive.*

   Precision= True positive/ (True positives + False positives)

3. *Recall: - measures the proportion of true positive among the all-actual positive instances.*

   Recall=True positive/ (True positive+ False Negative)

4. *F1 -score: - a measure that combines precision and recall to provide and overall measure of a model's performance.*

3.2 procedure

3.2.1 any further processing?

Further data processing such as to use standardscaler is to improve the performance of the machine learning models. Many machine learning algorithms such as regression are sensitive to the scale of input feature. when some features have a larger scale than other so it can lead to biased results and reduced performance. By StanderScaler, we ensure that all feature have similar scale and thus avoid this issue.
StandardScaler is typically used as a preprocessing step before training a machine learning model. It involves fitting the StandardScaler on the training data and then transforming both the training and test data using the same scaler. This ensures that the scaling is consistent across the training and test data.

The main use of StandardScaler is to standardize the features of a dataset to improve the performance of machine learning models, especially those that are sensitive to the scale of the input features.

3.2.2 Models used:

There are several models used in machine learning classification, but for this use case we have label data i.e. target column contains grade is high ,low, medium. Some of the popular one is SVC, decision tree classifier, K-Nearest Neighbor and naïve bayes. and check which model gives best suitable accuracy and which model we can deploy further.

4.0 Discussion:

Justification of four algorithms:

Supervised machine learning algorithms are a type of machine learning algorithm that learns to make predictions based on labeled training data. In supervised learning, the algorithm is trained using a set of input-output pairs, where the input is the data and the output is the corresponding label or target value that we want the algorithm to predict. The goal of the algorithm is to learn a mapping between the inputs and outputs, so that it can make accurate predictions on new, unseen data.

There are many different types of supervised learning algorithms, including:

Regression: Regression algorithms are used to predict a continuous value, such as a price or a temperature.

Classification: Classification algorithms are used to predict a categorical value, such as whether an email is spam or not.

4.1 Decision trees:

Decision trees are a type of algorithm that make predictions by partitioning the data into smaller subsets based on a set of rules or conditions.

A decision tree is a type of supervised learning algorithm used in machine learning for both classification and regression tasks. It is a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Each internal node of the tree represents a decision on a feature, and each leaf node represents a class label or a numerical value. The tree is built by recursively partitioning the feature space into smaller and smaller regions, based on the values of the features, until the regions are pure or meet some other stopping criterion.

In decision tree classification, the goal is to classify a new observation by traversing the decision tree based on the observed values of the features, starting from the root node and proceeding down the tree until a leaf node is reached. In decision tree regression, the goal is to predict a numerical value by traversing the decision tree and computing the value at the leaf node reached by the observation.

The decision tree algorithm works by selecting the feature that provides the most information gain, or reduction in impurity, at each node. Information gain is typically measured by entropy or Gini impurity, which are measures of the degree of randomness or uncertainty in the class distribution at a given node. The algorithm stops when all the observations in a node belong to the same class, or when some other stopping criterion is met, such as a maximum depth or a minimum number of observations per node.

4.2 Support Vector Machines (SVMs):

SVMs are a type of algorithm that learn to separate the data into different classes based on a hyperplane that maximizes the margin between the classes.in our dataset we use SVC. SVC (Support Vector Classification) is a type of supervised learning algorithm used in machine learning for binary and multi-class classification. It is based on the concept of finding the optimal hyperplane that separates the different classes in a high-dimensional space.

In SVC, the training data is used to find the optimal hyperplane that maximizes the margin, which is the distance between the hyperplane and the closest data points of each class. The hyperplane is chosen such that it maximizes the margin and correctly classifies the training data.

In cases where the data is not linearly separable, SVC uses a kernel function to map the input data into a higher-dimensional space where it is more likely to be linearly separable. The most common kernel functions used in SVC are the linear, polynomial, and radial basis function (RBF) kernels.

The hyperparameters of the SVC algorithm, such as the regularization parameter (C) and the kernel parameters, can be tuned using cross-validation or other model selection techniques to optimize the performance of the algorithm on the given dataset.

SVC has been shown to be effective in a wide range of applications, including image classification, text classification, and bioinformatics.

4.3 Naïve bayes:

Naive Bayes is a type of probabilistic classification algorithm used in machine learning. It is based on Bayes' theorem, which describes the probability of an event occurring based on prior knowledge of conditions that might be related to the event. Naive Bayes is called "naive" because it assumes that the features used for classification are independent of each other, which is often not the case in practice.

In Naive Bayes, the algorithm builds a probabilistic model of the training data, assuming that the distribution of each feature is independent of the other features. Then, for a new observation, the algorithm calculates the probability of each class given the observed values of the features, using Bayes' theorem. The class with the highest probability is assigned as the predicted class for the new observation.

4.4 k-nearest neighbor:

In KNN classification, the K nearest neighbors is found based on a distance metric, such as Euclidean or Manhattan distance, and the class of the new observation is assigned based on the majority class of the K neighbors. In KNN regression, the K nearest neighbors is used to calculate the average or median value, which is assigned as the predicted value for the new observation.

The choice of the value of K is an important parameter in KNN. A larger value of K makes the algorithm more robust to outliers and noise, but can lead to overfitting in some cases. A smaller value of K makes the algorithm more sensitive to noise and may result in overfitting. The optimal value of K can be determined through cross-validation or other model selection techniques

There are three main types of Naive Bayes classifiers: Gaussian Naive Bayes, Multinomial Naive Bayes, and Bernoulli Naive Bayes. Gaussian Naive Bayes is used for continuous numerical data, Multinomial Naive Bayes is used for discrete count data, such as text data, and Bernoulli Naive Bayes is used for binary data.

The Naive Bayes algorithm is computationally efficient and requires relatively small amounts of training data to achieve high accuracy. It has been used in a variety of applications, such as text classification, sentiment analysis, spam filtering, and medical diagnosis. However, its performance may suffer if the assumption of independence between features is strongly violated, or if the training data is imbalanced or noisy.

Result:-

Here the comparison of the algorithm that I used .

It can be seen that which algorithm is best suitable for the deployment. in our dataset we have decision tree algorithm which is not suitable because of overfitting and other KNN , Naïve bayes, SVM have accuracy as below.

classifier	class	labels	PREC	REC	F1-SCORE	SUPPORT	ACC
Decision tree algorithm	*Predict the milk quality using ml*	*0(high)* *1(low)* *2(medium)*	*0.00* *1.00* *0.62*	*0.00* *0.98* *1.00*	*0.00* *0.99* *0.77*	*76* *115* *127*	*100%*
K-Nearest neighbor	*Predict the milk quality using ml*	*0(high)* *1(low)* *2(medium)*	0.84 0.86 *0.89*	0.74 0.97 0.87	0.79 *0.91* *0.88*	86 102 130	*86.79%*
Support vector classifier	*Predict the milk quality using ml*	*0(high)* *1(low)* *2(medium)*	*0.71* *0.90* *0.95*	*0.92* *0.85* *0.83*	*0.80* *0.88* *0.89*	*76* *115* *127*	*85.84%*
Naïve bayes	*Predict the milk quality using ml*	*0(high)* *1(low)* *2(medium)*	*0.86* *0.94* *0.89*	*0.84* *0.94* *0.91*	*0.85* *0.94* *0.90*	*76* *115* *127*	*90.00%*

5. Conclusion:

From this entire algorithm model, we can conclude that with this new machine learning AI technology we can reduce human effort. an maximize aur business profit. Not only for this particular domain but also for every sector we can have like image and object recognition, we can use machine learning to analyze patterns and detect anomalies in financial transactions, which can help to identify and prevent fraudulent activity, healthcare analysis these are just few examples of the many applications of machine learning AI in various industries.

But there are some limitations also like algorithm more robust to outliers and noise so it gives accuracy more than 90% practically it is not good for because model is overfitted that's why we need to use multiple algorithms to check best suitable accuracy for our given dataset.

Overall, building a machine learning model to predict milk quality require careful selection and processing of data, a suitable machine learning algorithm, and careful evaluation and fine-tuning of the model. With proper implementation, such a model can provide valuable insights into the quality of milk products, allowing dairy farmers and milk processors to maintain high standards and improve their operations.

References:

1."Pattern Recognition and Machine Learning" by Christopher M. Bishop
"Machine Learning: A Probabilistic Perspective" by Kevin P. Murphy

2."Deep Learning" by Ian Goodfellow, Yoshua Bengio, and Aaron Courville

3."Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" by Aurélien Géron

4."Python Machine Learning" by Sebastian Raschka and Vahid Mirjalili
"The Hundred-Page Machine Learning Book" by Andriy Burkov

5."Introduction to Machine Learning with Python" by Andreas C. Müller and Sarah Guido

6"Reinforcement Learning: An Introduction" by Richard S. Sutton and Andrew G. Barto

7."Machine Learning Yearning" by Andrew Ng
"Bayesian Reasoning and Machine Learning" by David Barber

8.*dataset from (kaggle kernels output prena0808/predict-the-milk-quality-with-ml -p /path/to/dest)*