1. INTRODUCTION

1.1 OVERVIEW:

This project aims to use artificial intelligence and machine learning techniques to predict and analyze the milk price. The project will collect data on the bases of quantity, fat and solid non facts(SNF)to predict price.

Additionally, To predict milk prices accurately, relevant data needs to be collected This typically includes milk prices, consumption patterns, and other factors that may influence the milk price.

1.2 PROBLEM STATEMENT:

To develop an artificial intelligence and machine learning based predictive model that can accurately predict the milk price and identify the factors that contribute to predict the price.

1.3 EXISTING SYSTEMS:

There are several existing systems and platforms that provide data and analysis on the milk price. Machine learning algorithms, such as random forests, support vector machines, or gradient boosting machines, can be employed for milk price prediction. Regression techniques, such as linear regression or multiple regression, can be applied to predict milk prices based on various factors that influence the market. The models estimate the relationship between these variables and milk prices to make predictions.

However, these existing systems have limitations, including a lack of comprehensive data analysis and prediction models that accurately predict the milk Therefore, this project aims to develop a comprehensive and accurate predictive model using artificial intelligence and machine learning techniques to provide valuable insights.

1.4 PROPOSED SYSTEM:

The proposed system will focus on analyzing and predicting the milk price. The system will use a Linear Regression model to develop predictive price and various dependent variables will be used to improve the accuracy of the predictions.

To better understand the data, the system will use various types of graphs, including line plots, box plots and bar plots. These plots will help visualize the data and identify any trends, patterns, or outliers.

Additionally, Clean the data by handling missing values, outliers, and inconsistencies. Perform data transformations or scaling if necessary to ensure data quality and normalize the variables.

Overall, the proposed system aims to provide accurate predictions and valuable insights. These insights can predict the accurate value.

1.5 OBJECTIVES:

The main objectives of this project are to:

1. Develop a predictive model to get the accurate value.
2. Analyze the factors that contribute, to predict the price base on shifts,quantity,fat and snf(solid non fact).
3. Milk price prediction helps farmers, dairy industry stakeholders, and policymakers make informed decisions related to pricing.
4. Market strategies. Accurate predictions enable them to optimize their operations, plan production levels, adjust pricing strategies, and manage risk effectively.

1.6 OVERALL ARCHITECTURE:

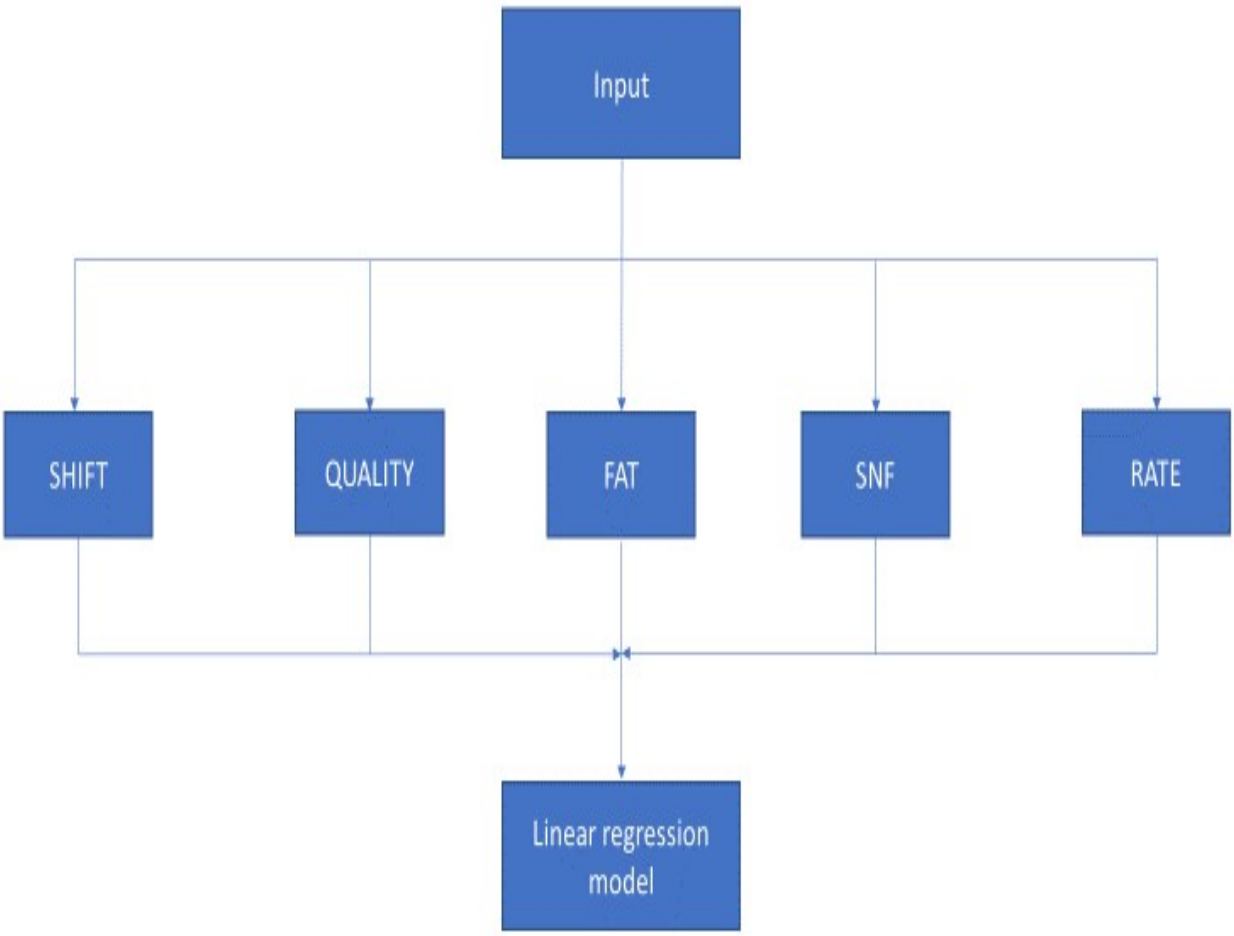

Fig:1.6.1 It is flowchart for our project model implementation

2. LITERATURE SURVEY

2.1 SURVEY DOCUMENTATION

In this section, we will document the literature survey we conducted related to the problem statement of predicting the milk price using artificial intelligence and machine learning techniques.

We conducted a comprehensive search of various databases, including IEEE Xplore, Google Scholar, and ScienceDirect. We used keywords such as milk price prediction," "machine learning," "artificial intelligence," and to identify relevant articles, conference papers, and preprints.

[1] Deluyker et al (1990): studied modeling daily milk yield in Holstein Cows using time series analysis. The study found that the ARIMA (0, 1, 1) model also known as the exponential smoothing model was fit to the daily yield from heifers and multifarious cows not treated fordisease and without missing milk weights Heman D.Lohano and Fateh M. Soomro (2006).

[2] Farhan Ahmed et al (2011): studied fore casting milk production in Pakistan. The aim of the study was toestimate the milk production in Pakistan for the period from 2010-11 to 2014-15. The study identified ARIMA (1,1,1) model is an appropriate model for estimation and forecasted that increasing trend of milk production in Pakistan. The study was estimated milk production of Pakistan increased to 47494.2 thousand tonnes in 2015-16 from 39650.1 thousand tonnes in 1990-91.

[3] Jai Sankar.T and Prabhakaran. R (2012): jointly studied forecasting milk production in Tamil Nadu. Thestudy found that ARIMA(AutoRegressive Integrated Moving Average) (1,1, 0) model was more suitedfor estimation. The study estimated that the milkproduction of Tamil Nadu would rise from 5.96 milliontonnes in 2008 to 7.15 million tonnes in 2015.

[4] Bjorn Gunnar Hansen (2014): studied differentmethods to forecast milk delivery to dairy: A comparisonfor forecasting. The study was chosen SARIMA model of(2, 1, 2)*(2, 1, 2) due to lowest AIC and BIC values. Themodel values of AIC, BIC, Variance and Likelihood were2016.31, 2051.06, 4932115 and -995.16 respectively.

[5] The Food Safety and Standards Authority of India (FSSAI) conducted the 'National Milk Safety and Quality Survey 2018' to address concerns about milk adulteration in India. The survey collected 6,432 milk samples from various sources across all states and union territories. Results revealed that only 12 samples were found to be adulterated, dispelling the perception of widespread adulteration. The FSSAI emphasized the need for continued efforts to ensure milk safety and quality in the country..

[6] Market Factors and External Data: Researchers have explored the influence of market factors, such as supply and demand dynamics, price indices of related commodities, milk inventory levels, and consumer behavior, on milk price prediction. Additionally, incorporating external data sources, such as social media sentiment analysis, news articles, or satellite imagery, has been explored to capture additional insights.

[7] Forecasting Horizon and Uncertainty Analysis: Studies have investigated different forecasting horizons, ranging from short-term to long-term predictions, and assessed the accuracy and stability of predictions over various time periods. Additionally, uncertainty analysis techniques, such as prediction intervals or probabilistic forecasting, have been explored to quantify and communicate prediction uncertainties.

[8] Comparative Studies and Evaluation Metrics: Researchers have compared different prediction models, techniques, or approaches to assess their performance on milk price prediction. Various evaluation metrics, such as mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), or

forecast skill scores, have been used to evaluate and compare the accuracy of prediction models.

[9] A comprehensive milk safety and quality survey was conducted by an independent agency on behalf of the Food Safety and Standards Authority of India (FSSAI). The survey collected 6,432 samples from various locations and found that 93% of the samples were safe for consumption, contradicting the perception of widespread milk adulteration. Only 12 samples were found to be adulterated, with low levels of contaminants unlikely to pose a serious health threat. However, the survey revealed concerns regarding aflatoxin M1 residues in processed milk and emphasized the need for measures to address this issue.

[10] A survey study was carried out around two large cities in Burkina Faso to contribute to the understanding of the situation of local milk production and milk processing. Twenty-two dairy farms associated with nine dairy processing units were selected for the study. Two separate questionnaires were used to investigate the prerequisites for animal production and milk processing.

[11] LocalCircles received over 45,000 responses from households located across all districts of Delhi, Noida, Ghaziabad, Gurugram and Faridabad. LocalCircles conducted a survey in Delhi and NCR to understand how households consume milk, their preferences, issues of milk quality and freshness. Nearly two in three consumers surveyed in Delhi-NCR said the milk they consume regularly is not pure, according to a survey conducted by the community platform LocalCircles.

[12] NASS publishes: a Milk Production report each month with total monthly milk production, milk produced per cow during the month, and average number of milk cows in the herd during the month for each of the 24 major milk producing States. U.S. level data are also estimated and published monthly. Each quarter, the Milk Production report includes quarterly production and inventory information by state for all states. The February report contains totals for the prior two years for milk cows and milk production. In addition, number of licensed dairy herds is published. All reports are released around the 20th of each month.

[13] Sundaram Satya (2013) through his market survey has reported that while India has the largest bovine population in the World, its cattle are the least productive, yielding almost five times less than the global average. Milk yield in India is 800-1000 litres an animal, per year, against the global average of 7000-8000 litres a year. The report assumes that things are going to change with the launch of the National Livestock Mission to attract investment and to enhance productivity. It informs that the central budget 2013-14 has made a provision of Rs. 3070 million for the mission. There is also a provision for increasing the availability of feed and fodder.

[14] Safa Abdelgadier Hassan et al (2018): Studied milk production forecasting in Khartoum state, Sudan.The study found that ARIMA (1, 0, 0) model was fit. Thestudy was estimated that the milk production of Sudanwould increase to 7.49 million tonnes in 2030-31 from6.81 million tonnes in 2017.

[15] SanaoKatkasame et al (1996): studied trend analysis on milk production traits in the dairy farming promotion organization of Thailand.The study found that the phenotypic trend on milk yield, milk fat and fat percentage obtained from a weighted regression methodwas 37.22 Kg, 1.32 Kg and -0.0104% respectively. Thegenetic trend on milk yield, milk fat and fat percentageobtained from a weighted regression method.

[16] Heman D.Lohano and Fateh M. Soomro (2006): studied jointly unit root test and forecast of milk production in Pakistan. The study was estimated by usingthe Ordinary Least Square (OLS) Method. The studyfound that the positive trend in milk production during theyear from 1971-72 to 2004-05. The forecasted milk production of Pakistan increased from 30.70 milliontonnes in 2005-06 to 44.33 for the year 2014-15.

3. DATA PRE-PROCESSING

3.1 DESCRIPTION OF DATASET

The milk_price-data.csv dataset is a collection of daily time-series data collected from dairy form .We took this data physical interaction with dairy farm holder and collected all the data the uses to predict the price of the milk.

The dataset contains around 6 variables, including shift,quantity,fat and snf(solid not fact). In addition to the basic information, there are also dependent variables that provide rate.

The primary target variable is the price of milk, which represents the market price at a specific time. It is usually measured in units. production data related to milk is often included in the dataset. This information may include the quantity of milk.

Our own data set contains 100 rows and 6 columns or attributes where each attribute is an independent and one attribute is based on remaining all the attributes values based on all independent attributes we are going to predict the price.Each column names:- shift, quantity, fat and snf(solid not fact).

Quantity: The quantity variable refers to the volume or quantity of milk produced or sold. It is typically measured in liters or gallons. Quantity provides insights into the overall supply of milk and is an important determinant of milk prices.

SNF (Solid-Not-Fat): SNF is a parameter used to measure the non-fat solids content in milk. It includes proteins, lactose, minerals, and other substances present in milk, excluding fat. SNF is usually expressed as a percentage. SNF content affects milk quality and can have an impact on milk prices.

The fat: content of milk is another crucial variable in determining milk quality. It is usually measured as a percentage of the total milk volume and can range from 3.0% to 6.0% in cow's milk. Fat content affects milk's taste, texture, and nutritional value and can have an impact on milk price.

3.2 DATA CLEANING

The purpose of data cleaning is to prepare the dataset for analysis and modeling by identifying and correcting errors, inconsistencies, and inaccuracies. The data cleaning process involves several steps, including but not limited to:

1. Removing duplicate values: Duplicate values in the dataset can distort the results and lead to incorrect conclusions. Therefore, it is important to remove duplicate values before analysis.

2. Handling missing or null values: Null or missing values can be due to various reasons such as data entry errors, system failure, or non-response. Null values can adversely affect the analysis and may result in incorrect conclusions. Therefore, it is important to remove them from the dataset or impute them with reasonable values.

3. Handling outliers: Outliers are data points that are significantly different from the rest of the data. Outliers can arise due to various reasons such as data entry errors, measurement errors, or natural variability. Handling outliers involves identifying them and deciding whether to remove them or keep them in the dataset.

A clean dataset is necessary for accurate analysis and modeling, which are essential for developing robust AIML solutions. Data cleaning is often an iterative process that involves multiple rounds of cleaning, analysis, and modeling to ensure that the dataset is suitable for the specific AIML task at hand.

We have removed the duplicate as well as null values present in the current dataset by using necessary commands to make the dataset clean and easier to access for better acquiring of the data.

Data Sampling: In some cases, the dataset may contain a large number of records, making it computationally expensive to work with. In such cases, data sampling techniques can be employed to select a representative subset of the data for analysis and modeling.

3.3 DATA VISUALIZATION

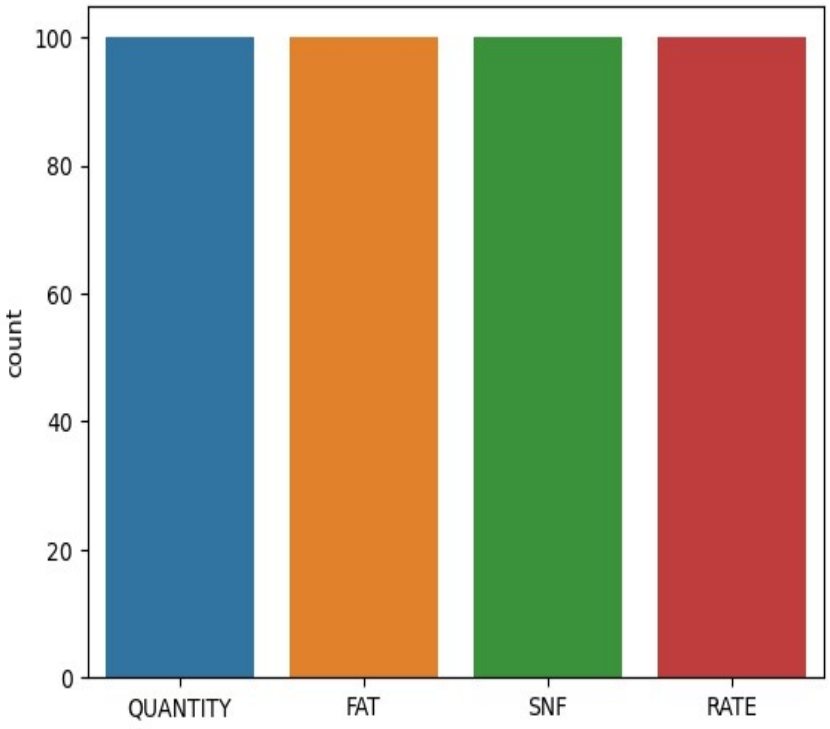

Fig:3.3.1

The countplot is used to represent the occurrence(counts) of the observation present in the variable. It uses the concept of a bar chart for the visual depiction.

The above bar count plot represents the number of values present in the each attribute (quantity, fat, and snf (solid not fact).From the above countplot it seen that all the variables has almost same 100 rows.

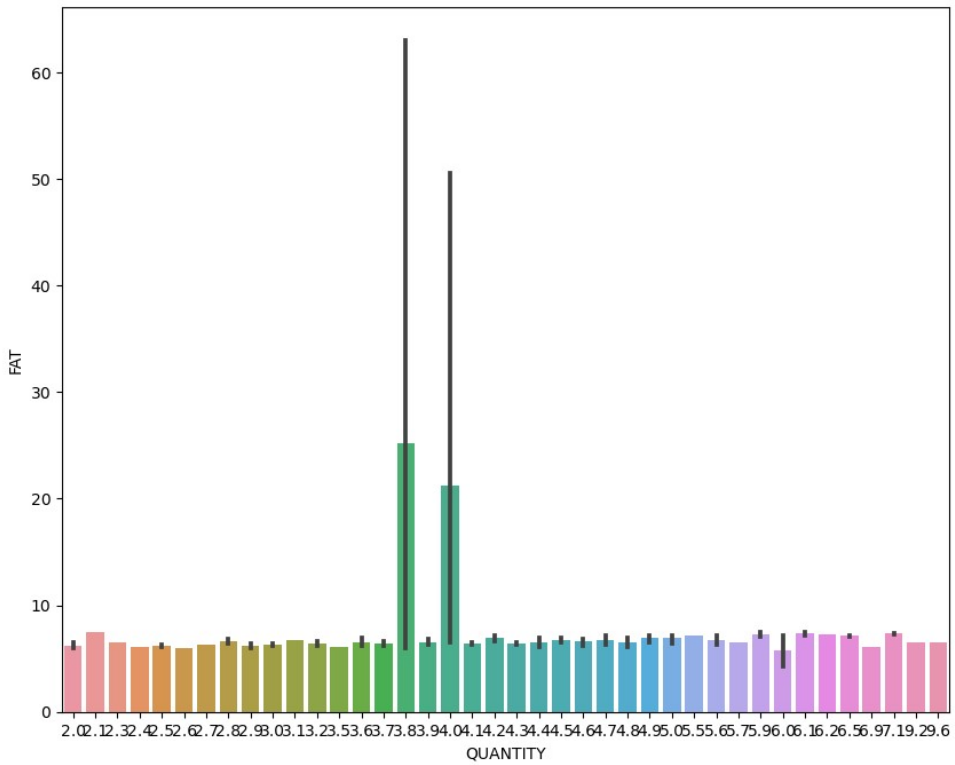

Fig:3.3.2

The above bar plot or bar graph is a chart or graph that with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a column chart.it represents fats and quantity of the milk.

From above bar plot shows for how much quantity of milk contains how much fat where quantity is taken on X_axis and fat is taken on Y-axis. Quantity: The quantity variable represents the amount of milk produced or sold. In a bar plot, you can display the different levels or categories of milk quantity on the x-axis, and the height of each bar represents the corresponding quantity value.

SNF (Solids-Not-Fat): SNF refers to the non-fat solids present in milk, such as proteins, lactose, and minerals. It can be represented in the bar plot by displaying different categories or levels of SNF on the x-axis, with the height of each bar indicating the SNF value for that category.

By using a bar plot to represent quantity, fat, SNF, and rate, you can compare the values of these variables visually. The bar plot allows you to easily observe the differences or similarities between categories or levels of the variables and provides a clear representation of their values.

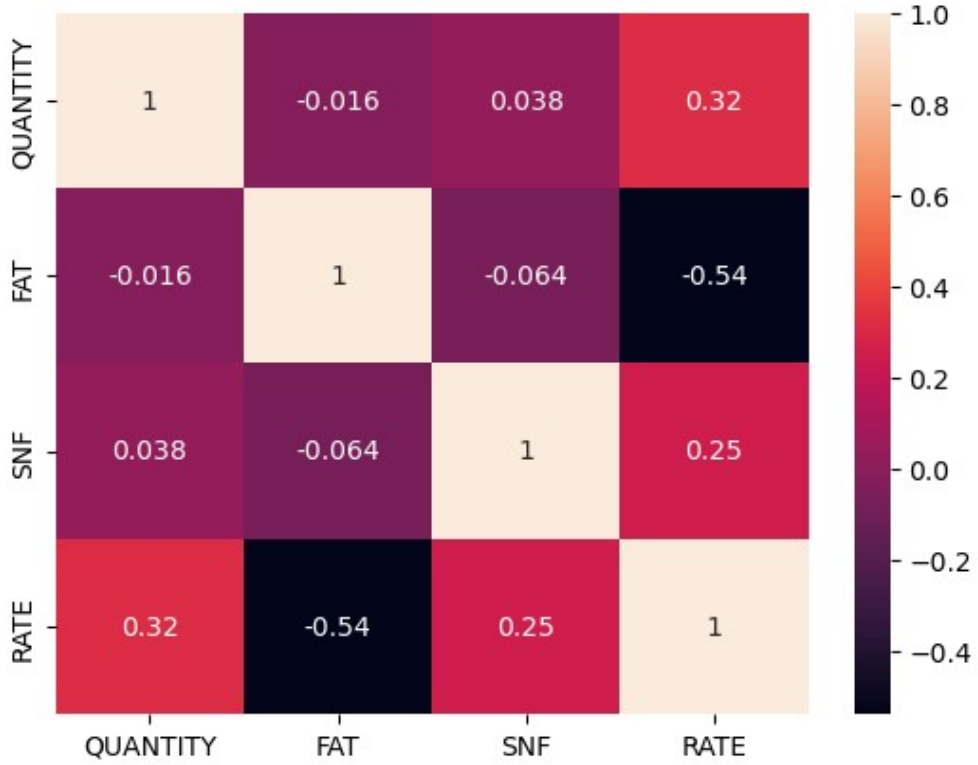

Fig:3.3.3

A heatmap depicts values for a main variable of interest across two axis variables as a grid of colored squares. The axis variables are divided into ranges like a bar chart or histogram, and each cell's color indicate the value of the main variable in the corresponding cell range.

The map contains variables:-quantity, fat, snf (solid not fact) and rate relates the varaibles with the corresponding variables.

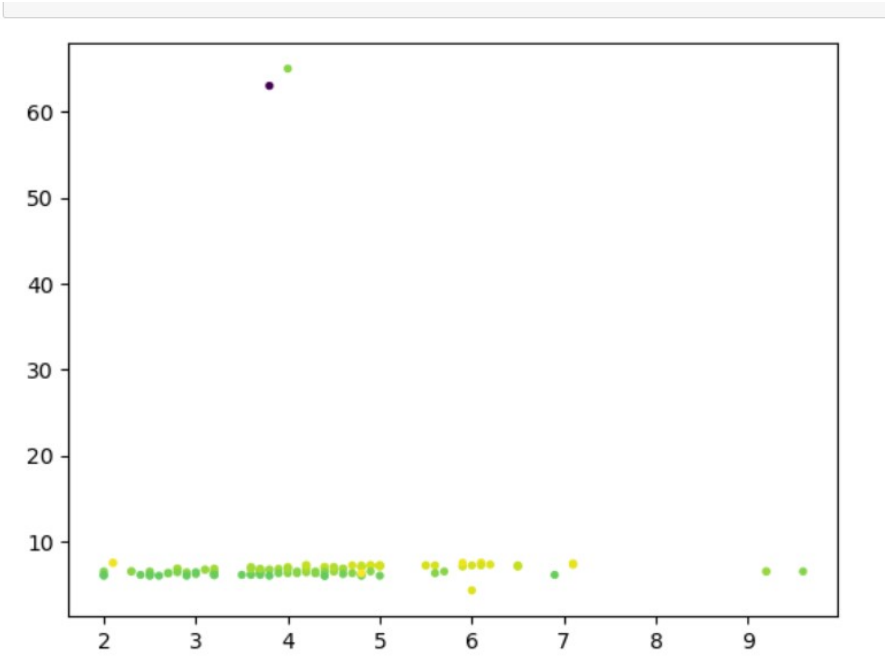

Fig:3.3.4

A scatter plot is a graphical representation that helps visualize the relationship between two variables. In the context of milk price prediction with the variables quantity, fat, SNF (solids-not-fat), and rate, a scatter plot can provide insights into the correlation or pattern between these variables and the milk price (rate). Here's a description of each variable and how they can be depicted in a scatter plot:

By plotting quantity, fat, and SNF against the milk price (rate) in a scatter plot, you can visually assess any potential relationships or patterns between these variables. The scatter plot allows you to observe if there is a linear, non-linear, or no apparent relationship between the independent variables (quantity, fat, SNF) and the dependent variable (rate).

Additionally, you can use different colors or markers to represent each variable in the scatter plot. For example, using blue dots for quantity, red dots for fat, green dots for SNF, and the position of the dots representing the corresponding rate value. This allows for easy identification and differentiation of variables in the plot.

.

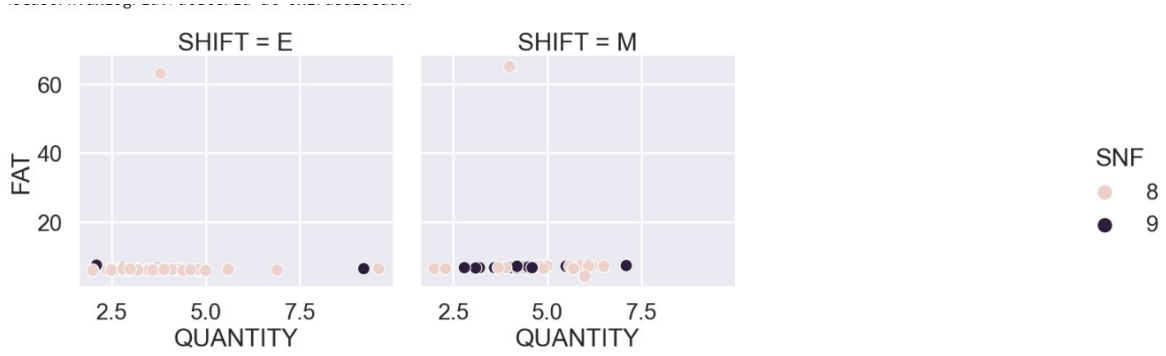

Fig:3.3.5

The relplot function is a part of the Seaborn library in Python, which provides a high-level interface for creating various statistical visualizations. It is particularly useful for creating relational plots that show the relationship between multiple variables. In the context of milk price prediction with the variables quantity, fat, SNF (solids-not-fat), and rate, the relplot function can be used to visualize the relationships between these variables.

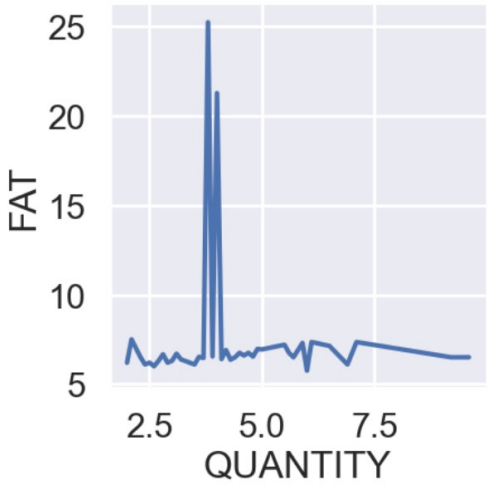

Fig:3.3.6

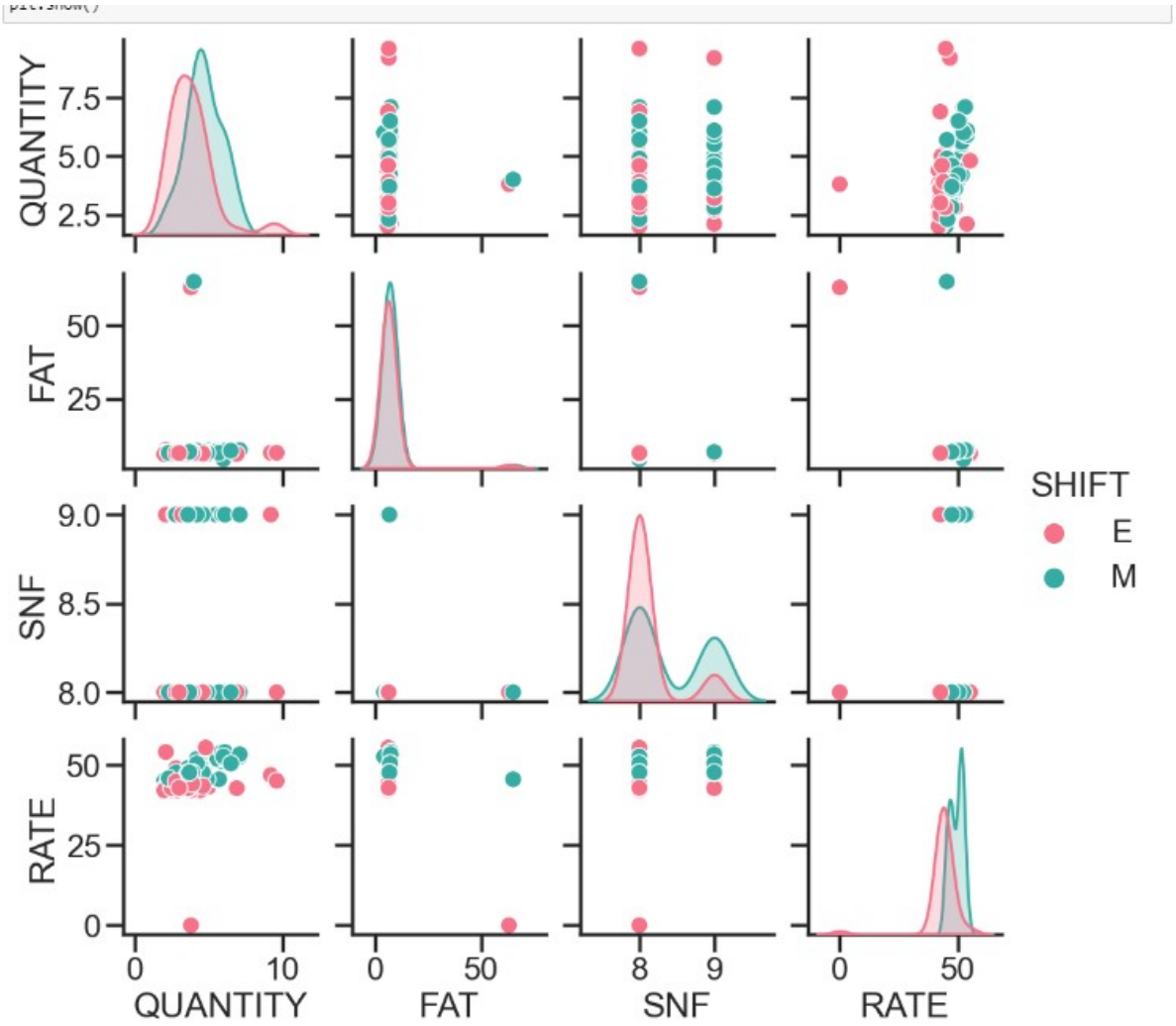

Fig:3.3.7

The pairplot function is a part of the Seaborn library in Python, which allows you to create a grid of scatter plots and histograms to visualize the pairwise relationships between multiple variables. In the context of milk price prediction with the variables quantity, fat, SNF (solids-not-fat), and rate, the pairplot function can be used to explore the relationships between these variables.

The pairplot function is particularly useful for exploratory data analysis, as it provides a comprehensive overview of the relationships between multiple variables in a concise visual format. It helps in identifying potential predictors or

understanding the interdependencies between the variables in milk price prediction .

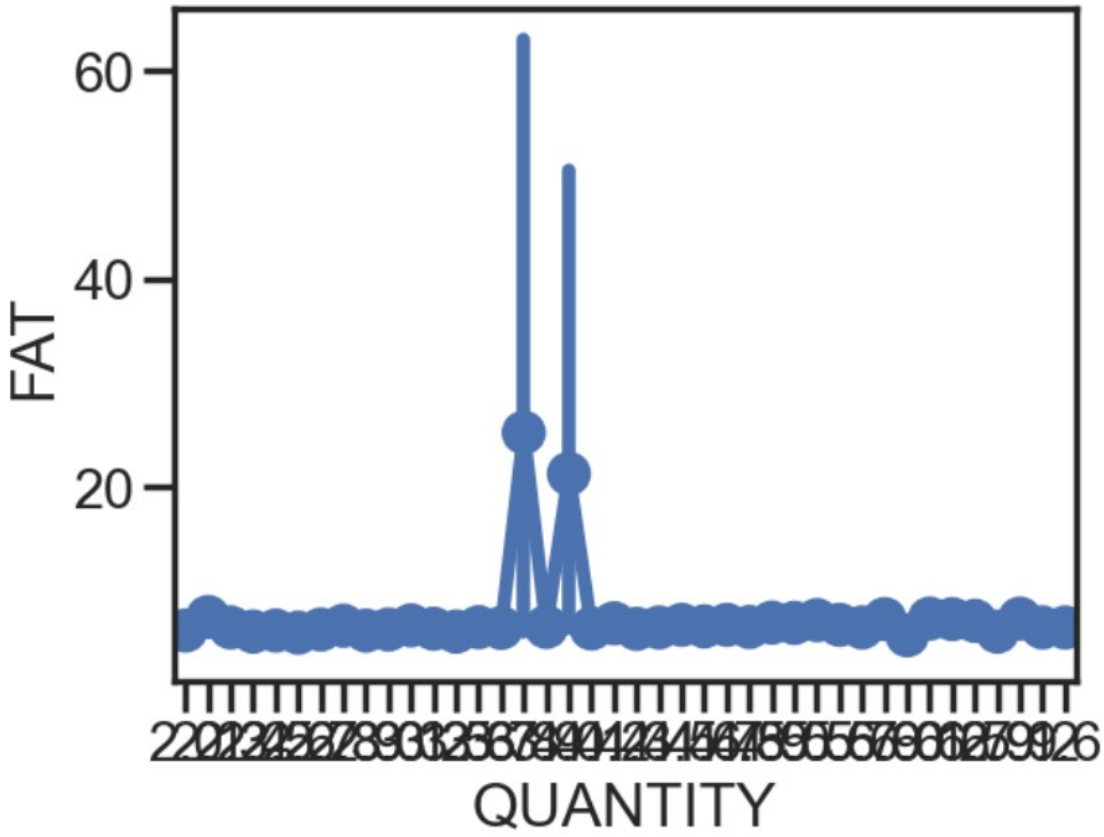

Fig:3.3.8

The pointplot function is a part of the Seaborn library in Python, which allows you to create a point plot to visualize the relationship between two categorical variables and a numeric variable. In the context of milk price prediction with the variables quantity and fat, the pointplot function can be used to explore the relationship between these variables.

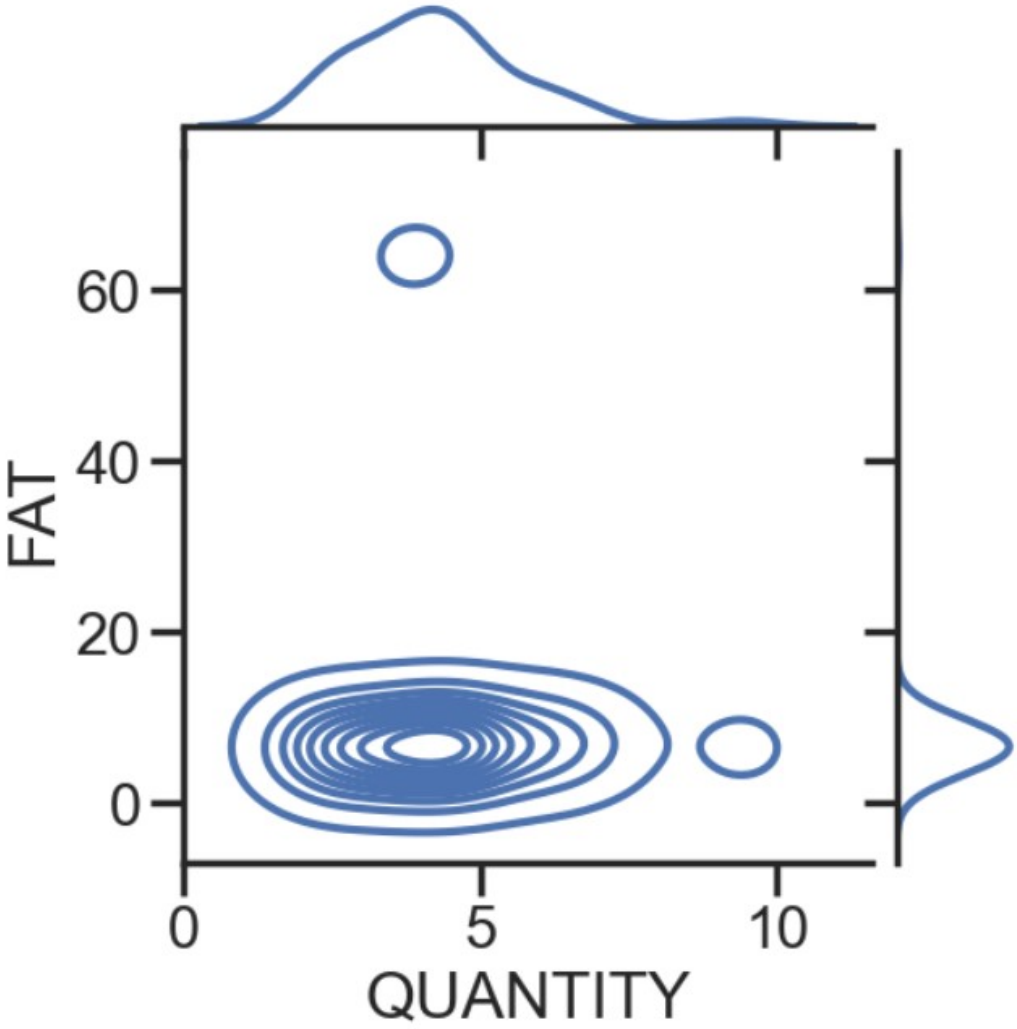

Fig:3.3.9

jointplot is a type of data visualization that allows you to explore the relationship between two variables by displaying their joint distribution and the individual distributions of each variable. In the context of milk price prediction, the variables of interest are quantity and fat content.

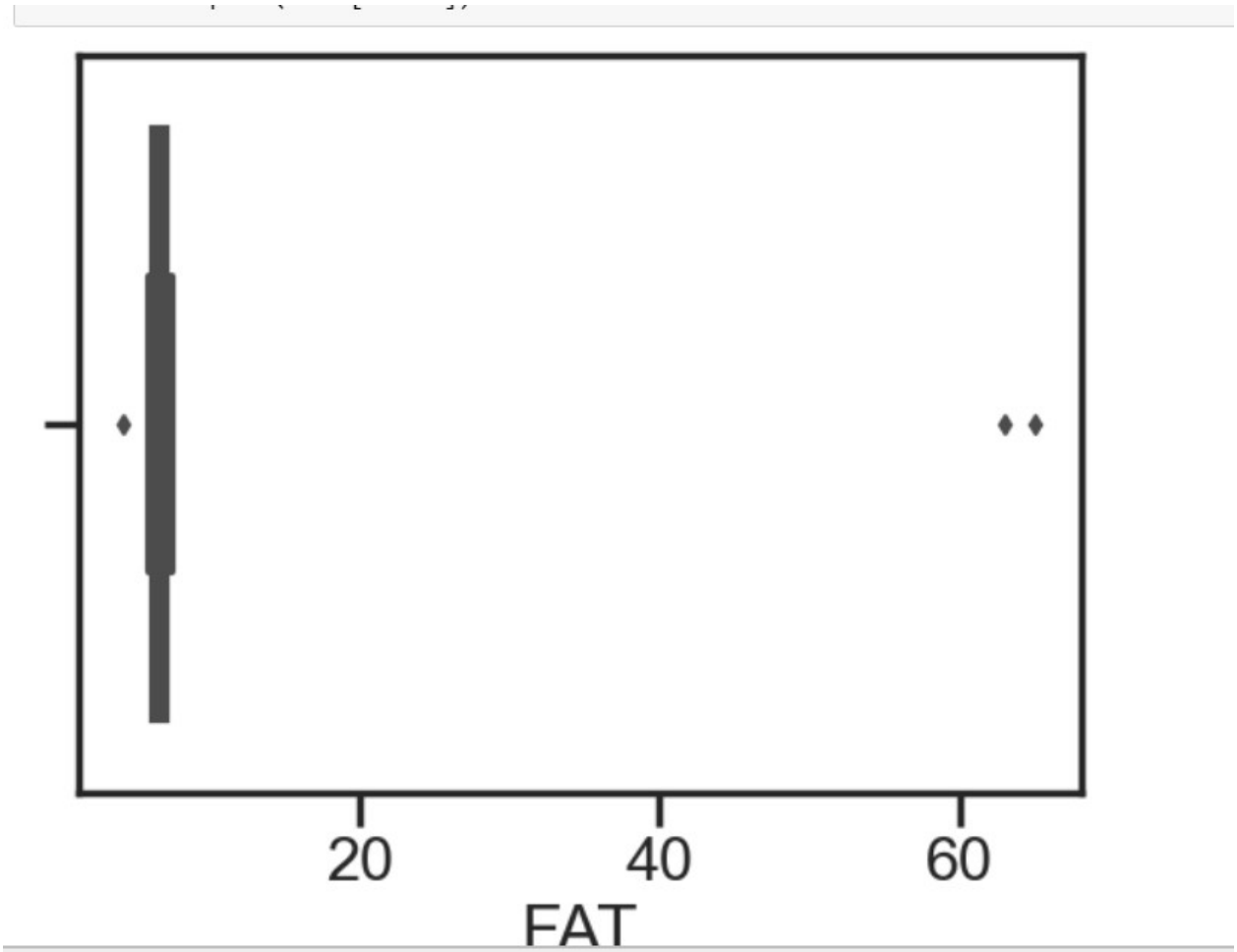

Fig:3.3.10

A boxplot, also known as a box-and-whisker plot, is a type of data visualization that provides a summary of the distribution of a variable. It is particularly useful when comparing distributions across different categories or groups. In the context of milk price prediction, the variables of interest are quantity.

utliers, which are data points that fall significantly outside the typical range, are usually represented as individual points.

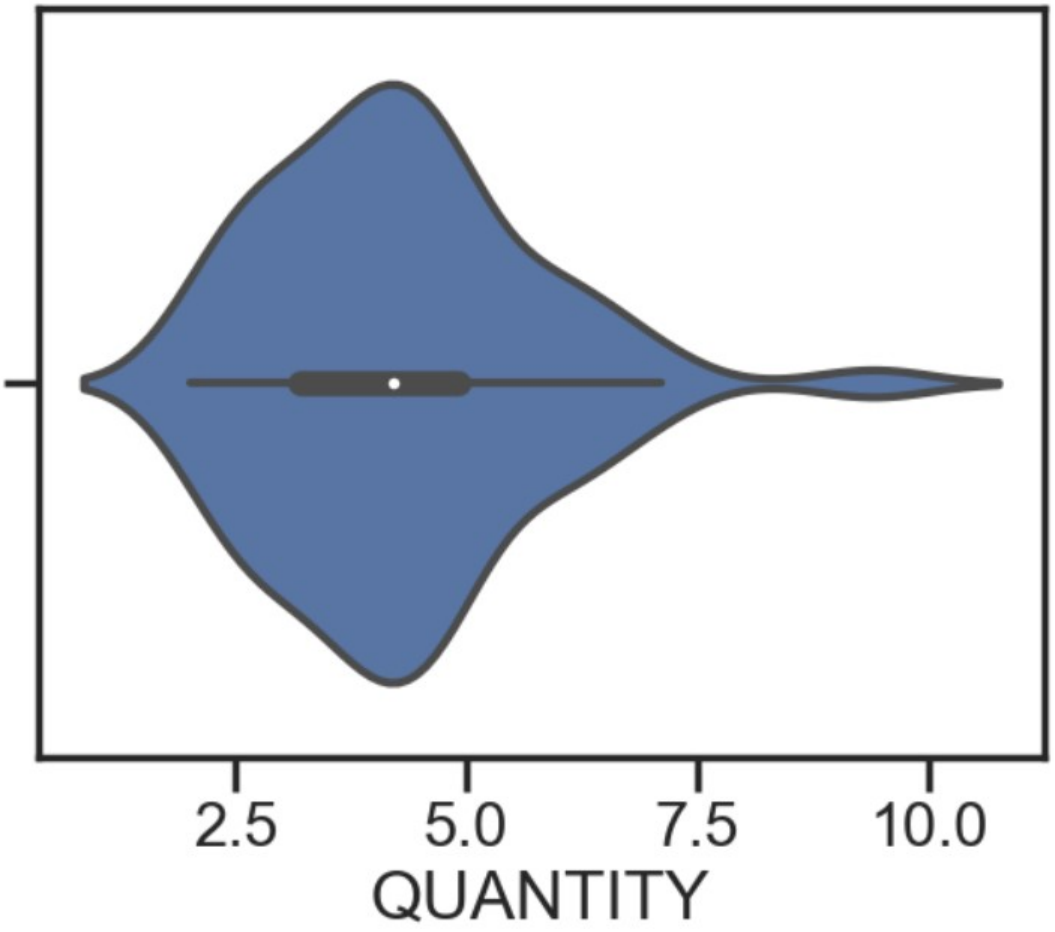

Fig:3.3.11

density plot to provide insights into the distribution and density of a variable across different categories or groups. In the context of milk price prediction, the variables of interest are quantity and fat.

Density Plot: Inside each violin, a kernel density plot is displayed, representing the estimated probability density of the data. This plot provides information about the concentration of milk prices within each category.

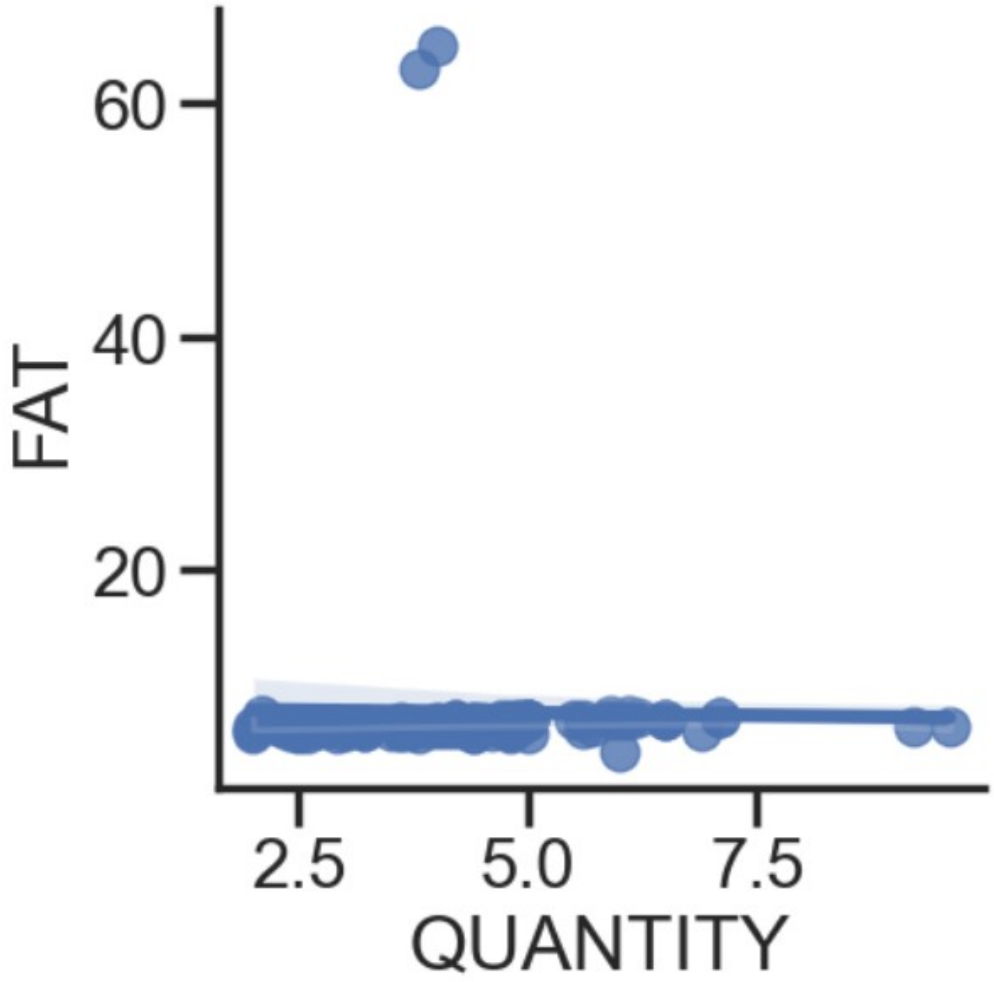

Fig:3.3.12

The fat content is plotted on the x-axis, the milk price on the y-axis, and the quality is represented by the color of the data points. A higher quality is indicated by warmer colors (reds), while lower quality is indicated by cooler colors (blues). The scatter plot allows you to visualize the relationship between milk price, fat content, and quality.

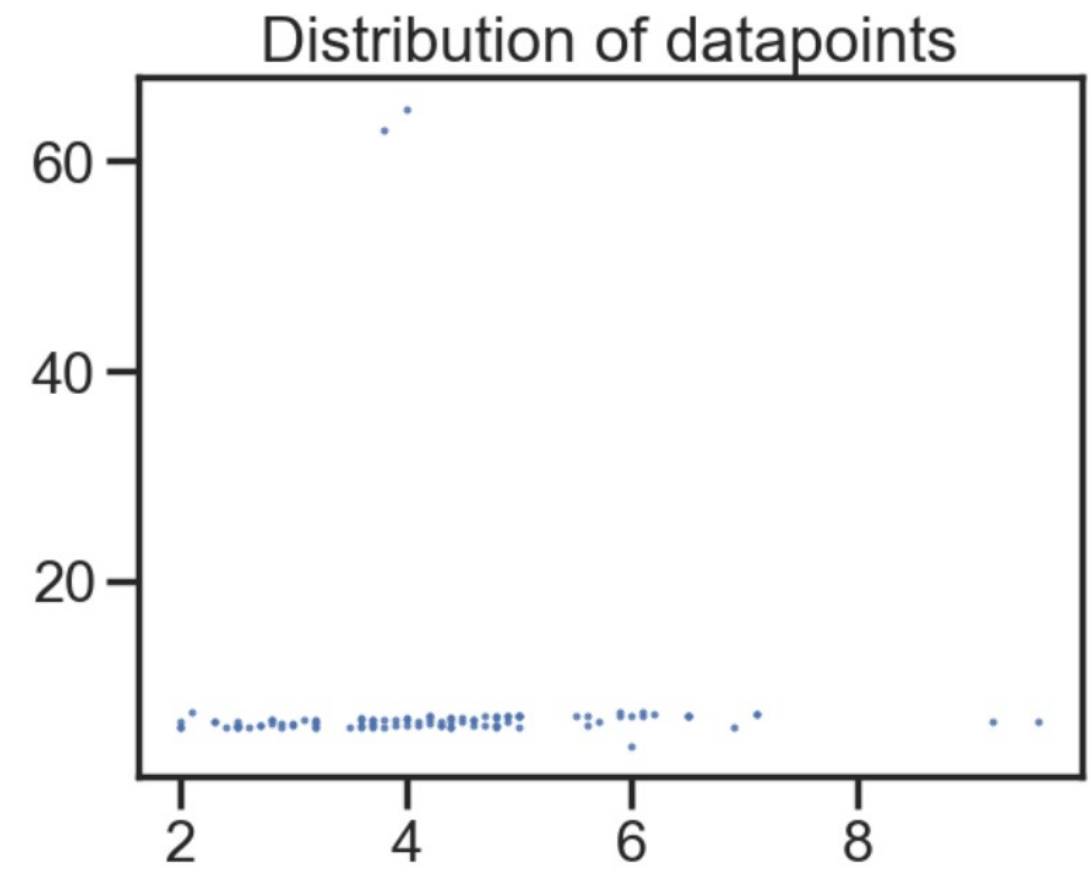

Fig:3.3.13

A scatter plot is a type of visualization that is useful for understanding the relationship between two continuous variables. In the case of milk price prediction, if you want to examine the relationship between milk price, fat content, and quality, a scatter plot can help you visualize the data points and identify any patterns or trends.

4. METHODOLOGY

4.1 PROCEDURE TO SLOVE THE GIVEN PROBLEM.

Linear Regression Model is used for our project in case of predicting the price. Linear Regression is an effective machine learning algorithm which there are one or multiple independent variables. These variables affect the dependent variable.

The goal of the linear regression model is to find the best-fit line that explains the relationship between the independent variables and dependent variables.

The linear regression model can be represented by the following equation:

$y = \beta0 + \beta1x1 + \beta2x2 + ... + \beta nxn$

where:

- y is the dependent variable (the variable being predicted)
- x1, x2, ..., xn are the independent variables
- β0, β1, β2, ..., βn are the coefficients (or parameters) of the model.

In this project, we have used the scikit-learn library to import the linear regression model. Before, implementing the linear regression model, we have cleaned the data to remove any null values and duplicated values.

After getting the train and test split of the data, we have implemented the linear regression model and then evaluated the model based on the various metrics like MAE, MSE, RMSE and R2 score.

At last, we have tested the current model by using making predictions by inputting some test data upon it. In case of there being too much error and too low

R2 Score, we have changed the percentage of train and test data to hyper tune the parameters, so that is more accurate.

4.2 MODEL ARCHITECTURE

Data collection: We physically visited the dairy farm and collected the dataset based on parameters that involved in predicting the price of the milk .We taken variables like quantity, fat and snf(solid not fact).

Data preprocessing: We have cleaned the data by dropping null values, selecting relevant independent variables (quantity,fat and solid non facts), and choosing a dependent variable (rate).

Train-test split: Split the data into training and testing sets using the train_test_split function from the sklearn library. We have chosen 95% of the data as train data and the remaining 5% for the test data.

Model selection and training: We have chosen a regression model, that is Linear Regression mode, and fit the model to the training data using the fit method. The dependent variable is rate and the independent variables are quantity fat and solid non facts(snf).

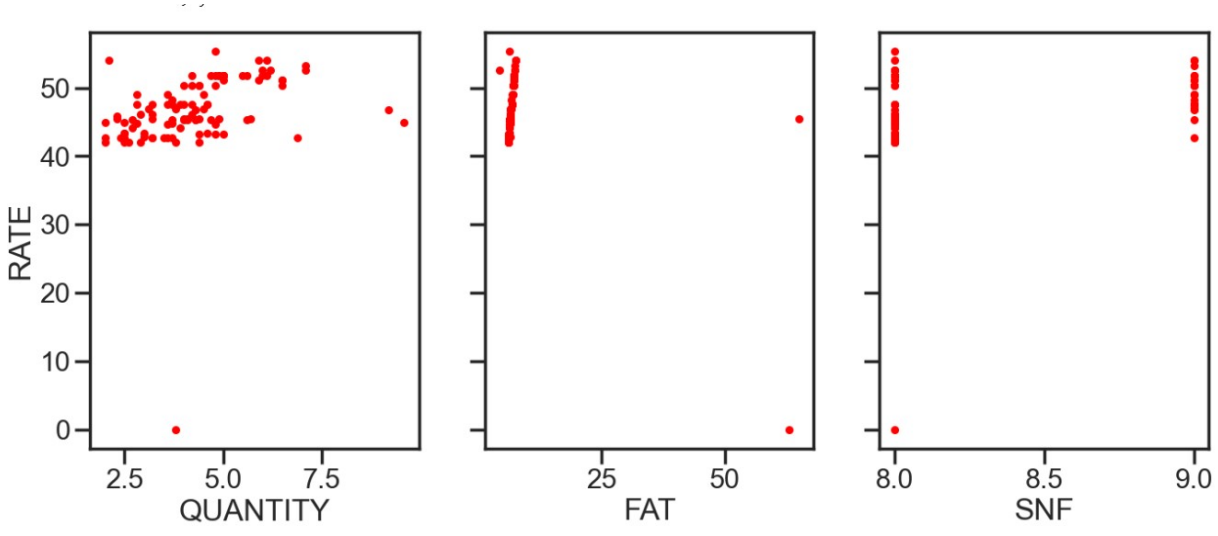

Fig:4.2.1

We have checked for the dependancy and relationship between the dependant and independent varibales and the relationship can be seen above.

Model evaluation: We have evaluated the model's performance on the testing set by calculating various metrics, such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R2 score.

```
Slope: [ 1.21056081 -0.38279373  2.73744134]
 Intercept: 21.815268571913077

Mean Absolute Error - MAE: 1.1085890947611745

Mean Square Error 2.00975111956486

Root Mean Square Error 1.417656911796666

R2 Score: 81.98786208374386 %
```

Fig:4.2.2

As shown in the above figure, we have obtained R2 Score as 81.98% which means that our model is accurate to a certain extent. The MAE, MSE and RMSE values have been calculated and noted down as shown above.

Prediction on new data:

We have tested the model on new data by making predictions using the predict method. We have created a new dataset of predicted variables and used the model to predict price.

Iteration and improvement:

We have analyzed the results and refined the model as needed, such as adding or removing variables, changing the model type, or adjusting parameters. We have adjusted the train and test split and have chosen the one with least error and better R2 score.

The model architecture can also be represented with the help of a block diagram.

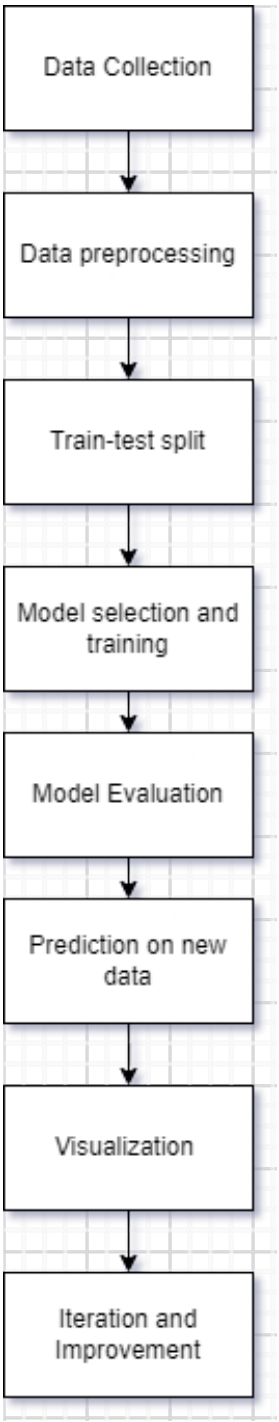

Fig:4.2.3

4.3 SOFTWARE DESCRIPTION

This project is developed using Jupyter Notebook, which is a popular web-based interactive development environment for creating and sharing data science projects. The code is written in Python programming language and uses several Python libraries, including Pandas, NumPy, Scikit-learn, Matplotlib, and Seaborn.

Pandas is a library that is used for data manipulation and analysis. It provides a set of data structures and functions to work with structured data, such as data frames and series. In this project, Pandas is used to load, clean, and manipulate the milk_price dataset.

NumPy is a library that is used for numerical computing with Python. It provides a powerful array processing capability and mathematical functions to work with large, multi-dimensional arrays and matrices. In this project, NumPy is used to perform numerical operations, such as calculating mean and standard deviation of the data.

Scikit-learn is a library that is used for machine learning tasks, such as building and evaluating models. It provides a wide range of tools for supervised and unsupervised learning, including regression, classification, clustering, and dimensionality reduction. In this project, Scikit-learn is used to build a linear regression model to predict the number of new price based on several input variables.

Matplotlib is a library that is used for creating visualizations and plots. It provides a set of functions to create various types of plots, such as line, bar, scatter, and histogram. In this project, Matplotlib is used to create scatter plots and regression lines to visualize the relationship between the input variables and the price prediction.

5. RESULTS AND DISCUSSION

The main objective of this project was to predict the milk price based on various factors such quantity, fat and solid non fact(snf).. The model used for this project was Linear Regression, which is a commonly used algorithm for predicting continuous variables.

The dataset we collected from one dairy farm from febuarary to April 2023. The dataset was preprocessed by dropping null values and splitting into training and testing sets. The training set was used to train the Linear Regression model, and the testing set was used to evaluate the model's performance.

The evaluation of the model was done using various metrics such as mean absolute error, mean squared error, root mean squared error, and R2 score. The model achieved an R2 score of 0.81, which indicates that 81.98% of the variance in the price can be explained by the independent variables used in the model.

Overall, the results of this project show that Linear Regression can be used to predict the price based on various factors. However, it is important to note that the model's accuracy may vary depending on the quality and quantity of data used in the model. Additionally, other machine learning algorithms can be explored to improve the model's performance.

	Actual value	predicted value
26	51.12	50.876714
86	50.40	48.827330
2	46.08	45.138714
55	42.00	44.565495
75	49.00	49.220208

Fig:5.1.1

6. CONCLUSION AND FUTURE SCOPE

Finally, this project proved how to apply machine learning approaches to forecast price based on characteristics such as quantity, fat and solid non facts(snf). The prediction model was built using the linear regression model, and several assessment criteria were employed to test the model's accuracy.

Based on the results, the model predicted price with a reasonable level of accuracy. The model, however, can be improved further by include additional features such as color , bread etc..

Furthermore, this project can be expanded by using more powerful machine learning models such as Random Forest, Gradient Boosting, or Neural Networks to improve prediction accuracy

7. REFERENCES

Milk price prediction:

We physically visited the farm and collected data set by taking parameters which makes to predict the price such as Quantity, Fat and solid non facts(SNF).

1.Pandas documentation (2021)

2.https://pandas.pydata.org/docs/

3.NumPy documentation. (2021)

4.https://numpy.org/doc/stable/

5.Scikit-learn documentation. (2021)

6.https://scikitlearn.org/stable/documentation.html

7.Matplotlib documentation. (2021)

8. https://matplotlib.org/stable/contents.html

9.Seaborn documentation. (2021)

10.https://seaborn.pydata.org/

11.Literature Survey Paper-I

IEEE Xplore Search Results

12. Literature Survey Paper-II

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3645145/