ABSTRACT :

The objective of this project is to establish a linear regression model that can estimate the impact of water quality factors on fish growth in aquaculture systems.

The project employs historical data collected from various aquaculture sites, encompassing water quality parameters such as temperature, dissolved oxygen levels, pH, and ammonia concentration. Through data analysis and feature engineering, the model establishes a relationship between these water quality indicators and fish growth.

By applying the linear regression model, fish farmers and aquaculture managers can obtain insights into the expected fish growth based on water quality conditions. This information enables proactive measures to maintain optimal water quality parameters, resulting in improved fish health and enhanced production efficiency.

The project’s results demonstrate the efficacy of linear regression as a predictive tool for water quality analysis in aquaculture. The simplicity and interpretability of the model make it accessible for practical implementation, facilitating informed decision-making in aquaculture management.

**Keywords :**

Water quality prediction, fish growth, linear regression, lasso regression, decision tree, aquaculture, predictive model.

INTRODUCTION :

Aquaculture plays a vital role in meeting the ever – increasing global demand for seafood. The success of aquaculture operations heavily relies on maintaining optimal water quality conditions to ensure the health and growth of fish species. Water quality parameters, such as temperature, turbidity, dissolved oxygen levels, pH, ammonia and nitrate significantly influence fish growth and overall aquaculture productivity.

To address this challenge, the use of predictive modelling techniques, specifically linear regression, lasso regression and decision tree has gained prominence in the field of aquaculture. Linear regression provides a valuable tool for predicting the relationship between water quality parameters and fish growth, enabling proactive management practices to optimize aquaculture systems.

The objective of this project is to develop a linear regression model that accurately predicts water quality parameters affecting fish growth. By establishing a quantitative relationship between water quality indicators and fish growth, aquaculture managers can make informed decisions regarding water quality management and create optimal conditions for fish growth.

By accurately predicting water quality parameters, aquaculture managers can take proactive measures to optimize water conditions, preventing adverse effects on fish health and maximizing production efficiency.

PROBLEM STATEMENT :

The growth and well-being of fish in aquaculture systems are heavily influenced by water quality parameters such as temperature, dissolved oxygen levels, pH, ammonia concentration, turbidity, and nitrate levels. However, traditional methods of monitoring and managing water quality in aquaculture often lack real-time capabilities and fail to capture subtle changes that can impact fish growth. This lead to suboptimal conditions, decreased productivity, and potential health issues for the fish.

The problem addressed in this project is the lack of an accurate and proactive approach to predict water quality parameters and their impact on fish growth in aquaculture systems. There is a need for a predictive model that can estimate the relationship between water quality indicators and fish growth, allowing aquaculture managers to make informed decisions and take proactive measures to maintain optimal water conditions.

Furthermore, the absence of a robust predictive model results in inefficiencies and increased costs for aquaculture operations. Without the ability to accurately forecast water quality parameters, aquaculture managers may face challenges in optimizing production processes, maintaining optimal fish health, and meeting the growing global demand for seafood in a sustainable manner.

Therefore, the problem statement for this project is to develop a linear regression model that accurately predicts the water quality parameters influencing fish growth in aquaculture systems. The model should enable proactive management practices by providing insights into the relationship between the water quality and the fish growth, allowing for timely adjustments in water conditions to optimize the fish health and the production efficiency.

**SOFTWARE DETAILS :**

* Jupyter Notebook
* Python

Special Packages Used :

* Pandas : Used for data manipulation and analysis
* NumPy : For numerical operations.
* Seaborn and Matplotlib : Used for data visualisations
* datetime : module that provides functions for working with

dates and time.

* StandardScaler and MinMaxScaler from sklearn.preprocessing:

Used for data scalling,ensuring that the input features are on a similar scale before training the linear regression model.

* train\_test\_split from sklearn.model\_selection:

Used to split the dataset into training and testing sets

* LinearRegression from sklearn.linear\_model:

Used to train the linear regression model on the training data and make predictions on the testing data.

* DecisionTreeRegressor from sklearn.tree:

Module for implementing decision tree-based models.

* mean\_squared\_error from sklearn.metrics:

Function is used to calculate the mean squared error between the actual fish lengths and the predicted fish lengths.

ABOUT THE DATA SET :

The datasets are generated from freshwater aquaponics catfish ponds. The IoT unit has six sensors (temperature, turbidity, dissolved oxygen, pH, ammonia, nitrate.

The attributes in the freshwater aquaponics catfish pond datasets are:

1. Date/Time
2. Temperature
3. Turbidity
4. Dissolved Oxygen(DO)
5. pH
6. Ammonia
7. Nitrate
8. Population
9. Length of Fish 10.

10.Weight of Fish

MODEL SELECTION AND TRAINING :

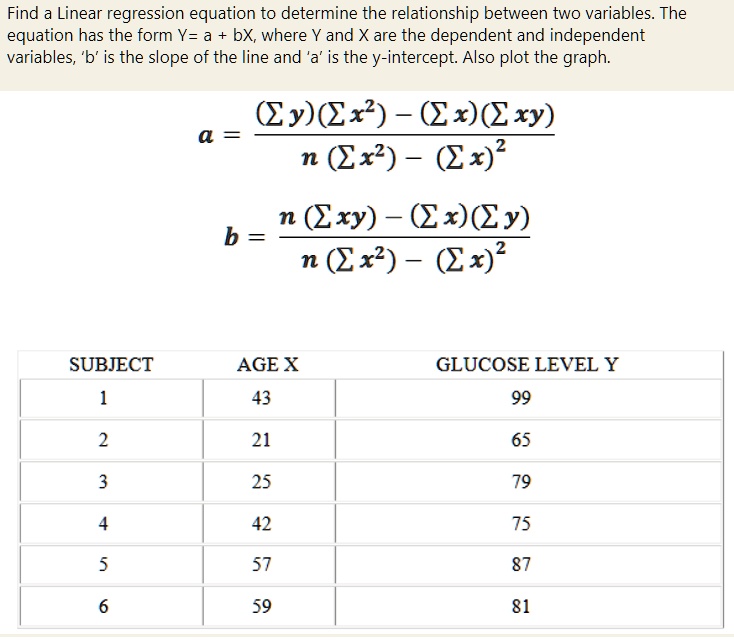
1. **Linear Regression** :

Linear Regression is a linear approach to modelling the relationship between the input features and the target variable. It assumes a linear relationship between the independent variables and the dependent variables. The model estimates the coefficients (slope and intercept) that define this linear relationship, allowing for the predictions of the target variable based on the input features. The mean squared error is used to evaluate how well the model’s predictions align with the actual values. Lower MSE values indicate a better fit of the model to the data.

A linear regression line quation is written in the form of :

y = a + bx

Now, here we need to find the value of the slope of the line, b, plotted in scatter plot and the intercept, a.



Here performs linear regression modelling and evaluates the model’s performance using the mean squared error ( MSE ):

* The LinearRegression class from scikit-learn. Linear regression is a supervised learning algorithm used for modelling the relationship between a dependent variable (y) and one or more independent variables (x)
* The linear regression model on the training data. The fit() method fits the model to the training data by estimating the coefficients that minimize the sum of squared residuals between the actual target values (y\_train) and the predicted values based on the input features (x\_train).
* The trained linear regression model to make predictions on the testing data (x\_test). The predict() method predicts the target variable (y) based on the input features (x\_test).

1. **Lasso regression :**

Lasso regression is a regularization technique used in the linear regression models to prevent the overfitting and improve the model performance. In the context of the fresh water aquaponics catfish pond project, Lasso regression can be applied to analyze the relationship between the water quality attributes (temperature, turbidity, dissolved oxygen, pH, ammonia, nitrate) and the target variable (e.g. fish growth, fish weight).

1. The equation of the lasso regression can be expressed as :

y = β0 + β1X1 + β2 X2 + …… + βpXp .

1. Where

y is the target variable or the dependent variable .

β0 is the y-intercept or the constant term.

β1, β2, β3…… βp are the coefficients of the independent variables x1,x2,…..xp.

x1,x2,…..xp are the independent variables or the features

1. The Lasso regression introduces a regularization term that adds a penalty to the loss function based on the absolute values of the coefficients :

**Loss Function = Sum of Squared Residuals + α \* ( sum of absolute values of coefficients)**

1. **Decision Tree :**

Decision Tree is a supervised machine learning Algorithm that can be applied to both classification and the regression tasks. It builds a tree – like model of decisions and their possible consequences based on the training data.

A decision tree works by partitioning the feature space based on the attribute values and their relationships with the target variable. It recursively splits the data into the subsets based on the selected features, aiming to minimize the MSE at each split.

The decision tree algorithm uses the following steps to construct the tree:

1. Select the best attribute:

The algorithm determines the attribute that best splits the data based on a criterion, such as maximizing the information gain or minimizing impurity.

1. Split the data:

The selected attribute is used to split the data into subsets. Each subset corresponds to a specific attribute value or range.

1. Create child nodes:

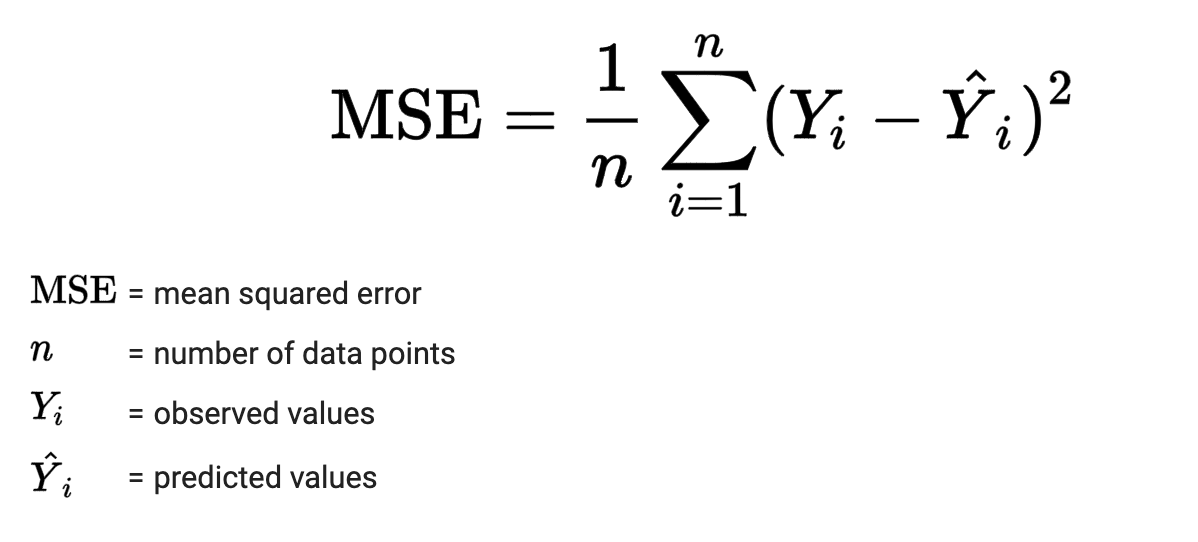
For each subset, a child node is created. The process of selecting the best attribute and splitting the data is recursively applied to each child node until a stopping criterion is met.

1. Assign a class or value to each leaf node:

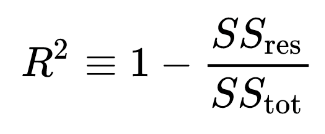
Once the tree is built, the leaf nodes are assigned a predicted class or a value. For regression problems, the predicted value can be the mean or median of the target variable within that leaf node.

MODEL EVALUATION :

* The mean square error or the MSE is evaluated by th formula:



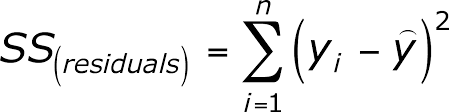
* The R2 Score is given by the formula :

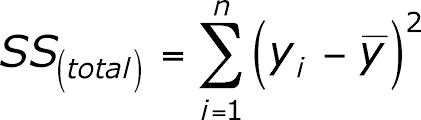


where,

SSres : Sum Squares due to residuals.

SStot : Total sum of squares.





**RESULTS AND DISCUSSIONS :**

1. **Data preparation** :

The Dataset initially contains the missing values and the outliers, which are handled in the code. Missing values in the ‘Ammonia(g/ml)’ column are replaced with the mean value. Outliers in the ‘Temperature(C)’, ‘PH’, and ‘Dissolved Oxygen(g/ml)’ columns are replaced with the previous values. Infinite values are replaced with the large value. Additionslly the dataset is split into features (X) and the target variable (y) for the training and the testing.

1. **From the Model Evaluation** :

* For the Linear Regression the MSE is 0.826 and the R2 Score is 0.9699.
* For the Lasso Regression the MSE is 0.8190 and the R2 Score is 0.9702
* For the Decision Tree Regressor the MSE is 0.00002 and the R2 Score is 0.9991

1. From Final Output :

* The x-axis represents the range of the temperature values, and the y-axis represents the frequency or the count of occurrences within the range.

It identifies the optimal temperature range for the fish growth. This range is typically associated with the most frequent or the highest bars in the histogram. It indicates the temperature level that are most suitable for the health and the growth of the fish.

The histogram shows that the most suitable temperature range lies between approximately 23 to 27 degrees Celsius.

* The x-axis represents the range of PH values and the y-axis represents the frequency within each range. The minimum and the maximum values on the x axis indicate the range of the pH values observed in the dataset.

It helps identifying the optimal pH range for the fish growth. The highest bars present on the histogram indicates the pH levels that are most suitable for the health and growth of the fish.

* The histogram presents a visual representation of the distribution of the dissolved oxygen levels in the water, The x axis represents the range of the dissolved oxygen levels, and the y axis represents the frequency within each range

It helps identify the specific ranges of dissolved oxygen levels that are critical for fish growth. These critical ranges might be associated with the optimal or suboptimal dissolved oxygen conditions for fish survival and growth. By analysing the histogram, you can determine the frequency or prevalence of dissolved oxygen levels within these critical ranges.

CONCLUSIONS :

1. The project aimed to predict the total length of fish in an IoT pond based on the water quality parameters. The analysis involved data preprocessing, exploratory data analysis, feature selection, and regression modelling using the 3 different algorithms: Linear Regression, Lasso Regression and Decision Tree Regression.
2. The results of this project indicate that the regression models were able to provide the predictions for the total length of the fish with varying levels of the accuracy. The MSE was used as a measure of the prediction error. The Decision Tree Regressor model achieved the lowest MSE compared to the rest two.
3. The findings of this project suggests that water quality parameters such as temperature, pH and the dissolved oxygen have a significant impact on the total on the total length of the fish in the pond.
4. However, it is important to note that the accuracy of the predictions can be further improved by incorporating additional features or exploring the different regression algorithms.
5. Overall project provides valuable insights into the relationship between the water quality parameters and the fish length in IoT pond. It demonstrates the potential of the regression modelling for predicting the water-quality related outcomes and can serve as a foundation for further research and the analysis in the field of aquatic ecosystems and the fish farming.