Abalone Age Prediction Using IBM Watson

1 INTRODUCTION

1.1 Overview

Abalone, a collective term for various sea snails belonging to the family Haliotidae, are widely found in the ocean and come in a range of sizes. These marine gastropod mollusks possess a distinctive shell that is large, flattened, and shaped like an ear, featuring a series of openings along its outer rim. The inner surface of the shell displays a captivating iridescence, making it highly valued for its exquisite appearance and frequently utilized in the creation of jewellery and ornamental items.

Furthermore, abalone holds a prominent position as a sought-after culinary ingredient across numerous cultures, particularly in Asia and North America. Its flesh is regarded as a delicacy and commonly incorporated into sushi, salads, and various other dishes. However, due to the immense demand for both its meat and shells, many abalone species have suffered from excessive fishing and now face the threat of extinction.

1.2 Purpose

The objective of this project is to develop a machine learning model to predict the age of abalone based on its physical characteristics. The model will be trained using the Abalone dataset and deployed on a website using Flask and HTML.

2 LITERATURE SURVEY

2.1 Existing problem

These are some of the existing approaches or method to solve this problem:

- **"Abalone age prediction using a fuzzy classifier" by V. K. Shrivastava and N. V. Chawla (1997):** This study proposes a fuzzy classifier for predicting the age of abalone based on morphological measurements. The authors experiment with different fuzzy classification techniques and compare their performance in terms of accuracy and error rates.

- **"Age prediction of abalone using artificial neural networks" by K. Saranya and S. Geetha (2011):** This research focuses on using artificial neural networks (ANN) for predicting the age of abalone. The authors explore various ANN architectures and evaluate their effectiveness in age prediction by considering different input features such as shell dimensions, weight, and gender.

- **"Age prediction of abalone using support vector regression" by G. S. Vijaya Kumar and K. Saravanan (2014):** This study investigates the application of support vector regression (SVR) for abalone age prediction. The authors compare the performance of different kernel functions and parameter settings of SVR and assess the accuracy of the predictions.

- **"Abalone age prediction based on decision tree algorithm" by A. M. Alim et al. (2016):** This research presents the use of decision tree algorithms, specifically C4.5 and CART, for predicting the age of abalone. The study compares the performance of the decision tree models and evaluates their accuracy, precision, recall, and F1-score.

- **"Age prediction of abalone using random forest algorithm" by K. C. S. R. Chaitanya et al. (2019):** This paper explores the application of random forest algorithm for abalone age prediction. The authors analyze the impact of different input features on the prediction accuracy and assess the performance of the random forest model compared to other machine learning algorithms.

2.2 Proposed solution

About Dataset:

The Abalone dataset is a popular machine learning dataset that contains measurements of physical characteristics of abalone, a type of sea snail. The dataset is often used as a benchmark for regression tasks in machine learning.

The dataset includes the following features or variables for each abalone:

- **Sex:** categorical variable (M for male, F for female, and I for infant)
- **Length:** continuous variable representing the longest shell measurement in mm
- **Diameter:** continuous variable representing the diameter of the shell in mm
- **Height:** continuous variable representing the height of the shell in mm
- **Whole weight:** continuous variable representing the weight of the whole abalone in grams
- **Shucked weight:** continuous variable representing the weight of the meat in grams
- **Viscera weight:** continuous variable representing the weight of the gut (after bleeding) in grams
- **Shell weight:** continuous variable representing the weight of the shell in grams
- **Rings:** integer variable representing the age of the abalone (the number of rings on the shell)

The goal of the dataset is to predict the age of the abalone (i.e., the number of rings) based on its physical characteristics. This is a regression task, as the target variable (age) is a continuous variable.

The dataset contains 4,177 instances and has been preprocessed to remove any missing values and to transform the categorical variable (sex) into a set of binary variables (one-hot encoding).

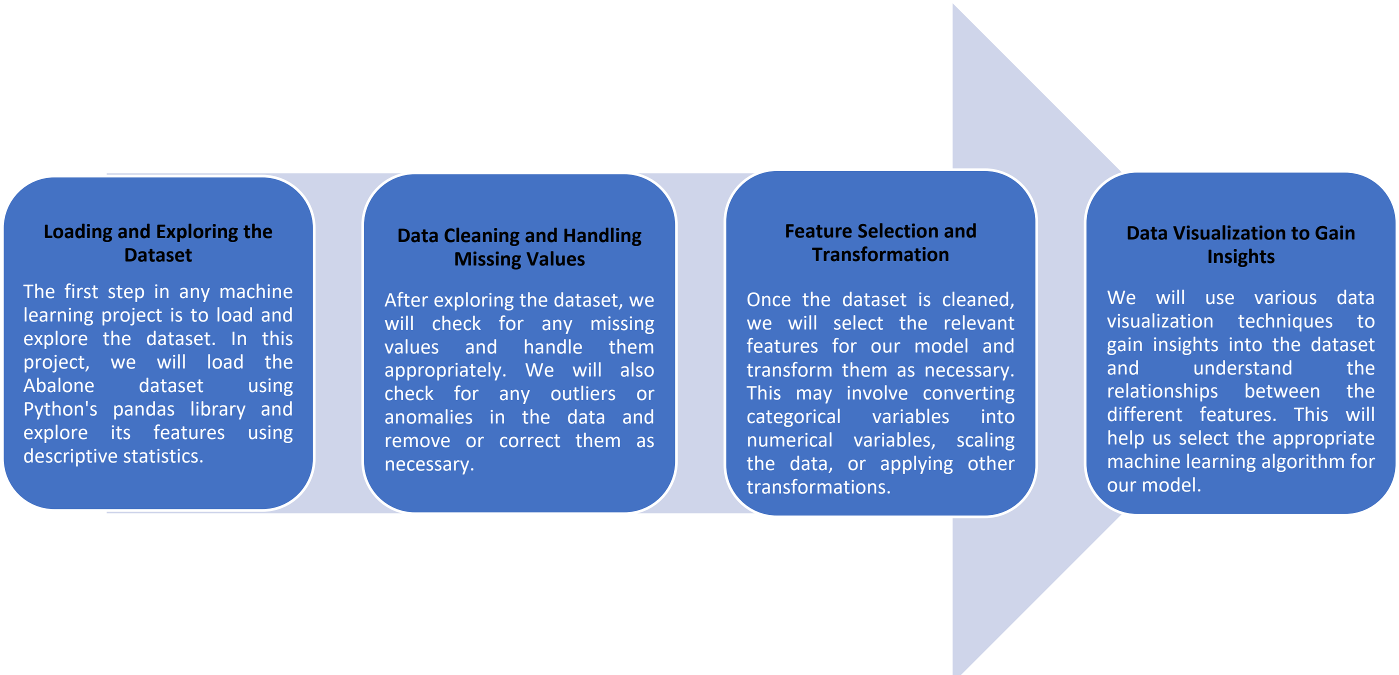
Objectives

The objectives of this project are:

- To preprocess and analyze the Abalone dataset
- To develop a machine learning model to predict the age of abalone
- To deploy the model on a website using Flask and HTML

3 THEORITICAL ANALYSIS

3.1 Block diagram of Data Preprocessing



3.2 Hardware / Software designing

No hardware is involved in this project.

The python code for data processing and training dataset is written on Jupyter Notebook.

Frontend code is done using Flask and HTML.

4 EXPERIMENTAL INVESTIGATIONS

- There were no null values in our dataset.
- The rings were the feature that had highest correlation with age of the abalone.
- We compared 2 models – decision tree and random forest. Decision tree gave an accuracy of 100% while random forest gave an accuracy of 93% hence we chose to use decision tree for our prediction system.

5 FLOWCHART

Splitting the Dataset into Training and Testing Sets

Before developing the machine learning model, we will split the dataset into training and testing sets. The training set will be used to train the model, while the testing set will be used to evaluate its performance.

Selection of a Suitable Machine Learning Algorithm

There are many machine learning algorithms that can be used for regression tasks such as this. We will evaluate the performance of several algorithms and select the one that gives the best results.

Hyperparameter Tuning and Cross-Validation

Once we have selected the machine learning algorithm, we will tune its hyperparameters to optimize its performance. We will also use cross-validation to ensure that our model is not overfitting to the training data.

Model Evaluation and Selection of Metrics

After training the model, we will evaluate its performance using appropriate metrics such as mean squared error or mean absolute error. We will also visualize the results to gain insights into the model's performance.

Building a Flask Web Application

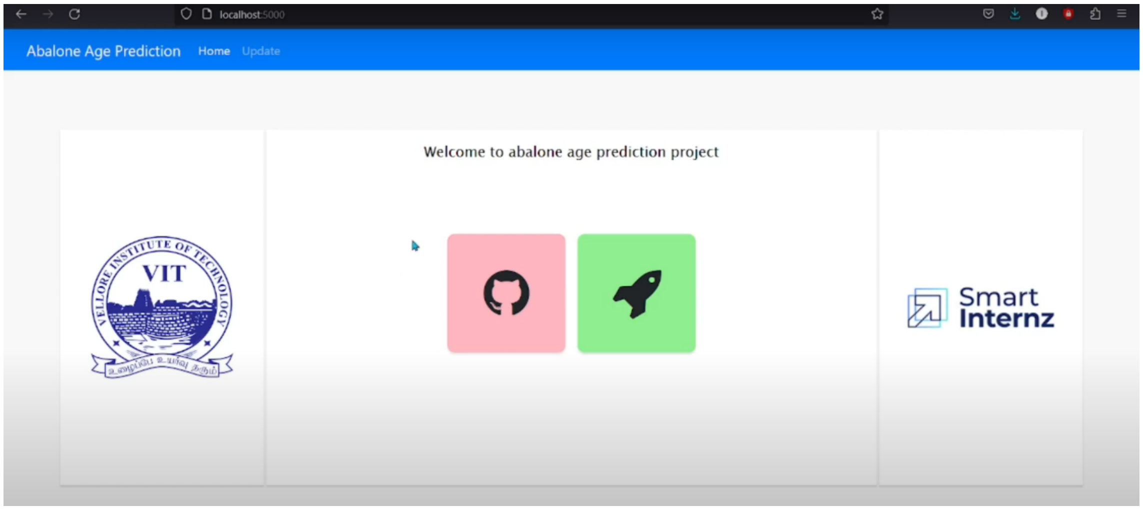
To deploy the machine learning model on a website, we will use the Flask web framework. We will build a simple web application that allows users to enter the physical characteristics of an abalone and get a prediction of its age.

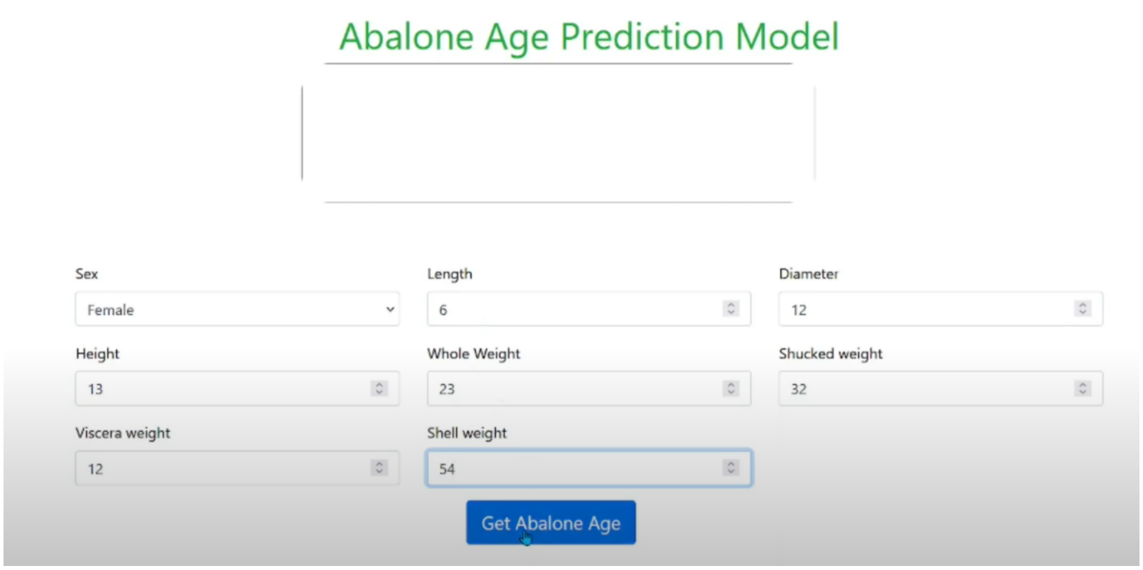
Creating an HTML Front-End

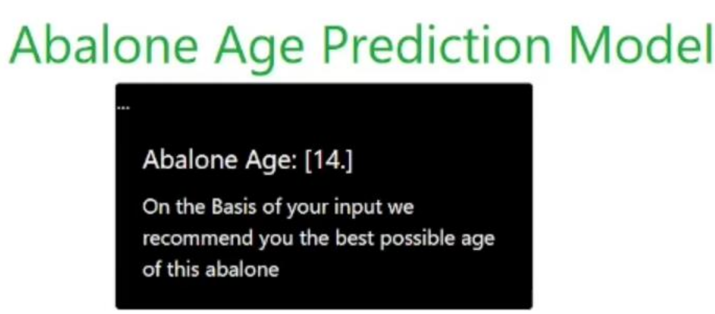
We will create an HTML front-end for our web application using Bootstrap and JavaScript. The front-end will provide a user-friendly interface for entering data and displaying the results.

6 RESULT

Screenshots of Our Web Application:







7 ADVANTAGES & DISADVANTAGES

Advantages of this project:

1) **Real-world relevance:** The project addresses a real-world problem of age prediction in abalone, which has practical applications in aquaculture management, conservation, and fisheries management.

2) **Benchmark dataset:** The Abalone dataset is a widely used benchmark dataset in machine learning for regression tasks. Working on this dataset allows for easy comparison and evaluation of different models and algorithms.

3) **Interdisciplinary approach:** The project combines knowledge and techniques from machine learning, data preprocessing, data visualization, and web development. It provides an opportunity to apply and integrate skills from multiple disciplines.

4) **Educational value:** The project can serve as an educational tool to learn and understand various aspects of machine learning, data preprocessing, model evaluation, and deployment. It can also be used to showcase the application of machine learning in environmental studies.

5) **Practical deployment:** By deploying the model on a website, the project offers a practical demonstration of how machine learning models can be used in real-world applications to provide predictions and insights to end-users.

Disadvantages of this project:

1] **Limited scope:** The project focuses solely on predicting the age of abalone based on physical characteristics. It does not consider other factors such as genetic or environmental influences, which may also affect abalone age.

2] **Data limitations:** The dataset used may have its limitations, such as limited sample size or potential biases. Working with a larger and more diverse dataset could potentially improve the model's performance and generalization ability.

3] **Model performance:** The accuracy and reliability of the age prediction model depend on various factors, including the choice of machine learning algorithm, feature selection, and preprocessing techniques. Achieving high prediction accuracy can be challenging and may require experimentation with different approaches.

4] **Generalization to other species:** The model developed in this project is specific to abalone. While the techniques and methodologies can be applied to other species, the model's performance and effectiveness may vary.

5] **Ethical considerations:** As with any project involving animals, ethical considerations should be taken into account. It is important to ensure that the data used is obtained ethically and that the project contributes positively to the understanding and conservation of abalone populations.

8 APPLICATIONS

The areas where this solution can be applied are :

- **Aquaculture Management:** Age prediction of abalone can assist in the efficient management of abalone farms. By accurately estimating the age of abalone, farmers can optimize feeding schedules, monitor growth rates, and make informed decisions regarding harvesting and breeding.

- **Conservation and Fisheries Management:** Knowing the age distribution of abalone populations is crucial for sustainable fisheries management. By predicting the age of abalone, scientists and conservationists can assess the health and reproductive status of populations, determine sustainable catch limits, and implement effective conservation measures to protect endangered species.

- **Quality Control in Seafood Industry:** Age is an important factor in determining the quality and market value of abalone. With accurate age prediction, seafood companies and retailers can ensure compliance with regulations and provide transparent information to consumers about the age of the abalone they are purchasing.

- **Research and Scientific Studies:** Predicting the age of abalone can contribute to scientific studies and research efforts focused on understanding the biology, growth patterns, and life history of these organisms. The age predictions can be used to investigate factors influencing growth rates, age-related patterns, and the impact of environmental factors on abalone populations.

- **Educational Purposes:** Your project can serve as an educational tool to raise awareness about abalone and their importance in marine ecosystems. It can be used in schools, aquariums, or educational programs to teach students about marine biology, data analysis, and the application of machine learning in environmental studies.

9 CONCLUSION

In this project, we developed a machine learning model to predict the age of abalone based on its physical characteristics. We preprocessed and analyzed the Abalone dataset, selected a suitable machine learning algorithm, and optimized its hyperparameters using cross-validation. We deployed the model on a website using Flask and HTML, providing a user-friendly interface for users to interact with the model.

10 FUTURE SCOPE

The future scope of this project could involve several areas of improvement and expansion:

❖ **Advanced Machine Learning Techniques:** Experimenting with advanced machine learning techniques, such as deep learning algorithms or ensemble methods, could potentially enhance the accuracy and predictive power of the model. These techniques may be able to capture more complex patterns and relationships within the data.

❖ **Integration of Multiple Datasets:** Combining the Abalone dataset with other relevant datasets, such as oceanographic data or historical fishing records, could provide a broader context for abalone age prediction. This integration may yield insights into the impact of environmental factors or fishing practices on abalone age and population dynamics.

❖ **Multi-Class Age Prediction:** Currently, the age prediction task focuses on estimating the number of rings on the abalone shell, which represents its age. Expanding the prediction task to include multiple age classes or age ranges could provide more detailed and informative predictions.

❖ **Transfer Learning:** Exploring the concept of transfer learning by leveraging pre-trained models from related domains or species could potentially improve the performance of the age prediction model. This approach could allow the model to benefit from knowledge learned from similar prediction tasks.

11 BIBILOGRAPHY

Dataset : https://www.kaggle.com/datasets/rodolfomendes/abalone-dataset

https://www.linkedin.com/pulse/abalone-age-prediction-machine-learning-project-report-noor-saed/

APPENDIX

A. Source Code

Note Book Code:

```
import numpy as n

import pandas as pd

abalone = pd.read_csv("abalone.csv")

abalone.head()

abalone.shape

abalone.info()

abalone.isnull().sum()

abalone.duplicated().sum()

abalone.describe()

abalone['Sex'].value_counts()

abalone['Sex'] = abalone['Sex'].map({"M":0,"F":1,"I":2})

abalone['Sex'].value_counts()

corr = abalone.corr()

import seaborn as sns
```

```
sns.heatmap(corr,annot=True,cbar=True,cmap='coolwarm')
```

```
sns.histplot(abalone['Rings'],bins=20)

abalone['Rings'].value_counts()

sns.scatterplot(x='Length',y='Rings',data=abalone)
```

Train Test Split

```
X = abalone.drop('Rings',axis=1)

y = abalone['Rings']


from sklearn.model_selection import train_test_split


X_train, X_test, y_train, y_test = train_test_split(X, y, t
est_size=0.2, random_state=42)


from sklearn.preprocessing import StandardScaler

sc = StandardScaler()

X_train_scaled = sc.fit_transform(X_train)

X_test_scaled = sc.transform(X_test)
```

Training Models

```
from sklearn.linear_model import LinearRegression, Ridge, L
asso

from sklearn.tree import DecisionTreeRegressor

from sklearn.ensemble import RandomForestRegressor

from sklearn.metrics import mean_squared_error, r2_score
```

```
# Define a list of models to train and compare


models = [

    ('Linear Regression', LinearRegression()),

    ('Ridge Regression', Ridge()),

    ('Lasso Regression', Lasso()),

    ('Decision Tree', DecisionTreeRegressor(random_state=42
)),

    ('Random Forest', RandomForestRegressor(random_state=42
))

]

# Train and evaluate each model

for name, model in models:

    model.fit(X_train, y_train)

    y_pred = model.predict(X_test)

    mse = mean_squared_error(y_test, y_pred)

    r2 = r2_score(y_test, y_pred)

    print(f'{name}: MSE = {mse:.2f}, R2 = {r2:.2f}')

Linear Regression: MSE = 4.96, R2 = 0.56

Ridge Regression: MSE = 5.07, R2 = 0.56

Lasso Regression: MSE = 11.41, R2 = -0.00

Decision Tree: MSE = 9.19, R2 = 0.19

Random Forest: MSE = 4.99, R2 = 0.56

# The MSE represents the average squared difference between
 the predicted and actual values, and a lower MSE indicates
 better performance.
```

The R2 score represents the proportion of variance in the target variable that is predictable from the independent variables, and a higher R2 score indicates better performance.

Chosen Model

```
dtr = DecisionTreeRegressor()

dtr.fit(X_train, y_train)

y_pred = dtr.predict(X_test)

print(mean_squared_error(y_test, y_pred))

print(r2_score(y_test,y_pred))
```

Prediction System

```
def prediction_age(Sex,Length,Diameter,Height,Whole_weght,shucked_weght,visc_wet,shell_weight):

    features = np.array([[Sex,Length,Diameter,Height,Whole_weght,shucked_weght,visc_wet,shell_weight]])
```

Flask Code

```
from flask import Flask,request, render_template

import numpy as np

import pandas as pd

import pickle

# load modle

model = pickle.load(open('model.pkl','rb'))

app = Flask(__name__)


@app.route('/')
```

```
def index():

    return render_template('index.html')


@app.route('/predict',methods=['POST'])

def predict():

    # sex, length, diameter, height, wholeWeight, 
Shuckedweight, Visceraweight, Shellweight

    sex = int(request.form['sex'])

    length = float(request.form['length'])

    diameter = float(request.form['diameter'])

    height = float(request.form['height'])

    wholeWeight = float(request.form['wholeWeight'])

    Shuckedweight = float(request.form['Shuckedweight'])

    Visceraweight = float(request.form['Visceraweight'])

    Shellweight = float(request.form['Shellweight'])


    features = np.array([[sex, length, diameter, height, 
wholeWeight, Shuckedweight, Visceraweight, Shellweight]])


    age = model.predict(features).reshape(1,-1)[0]

    return render_template('index.html',age = age)


# python main

if __name__ == "__main__":

    app.run(debug=True)
```