Nitte Meenakshi Institute of Technology

(AN AUTONOMOUS INSTITUTION AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM)

PB No. 6429, Yelahanka, Bangalore 560-064, Karnataka

Telephone: 080- 22167800, 22167860

Fax: 080 – 22167805

DATA MINING PROJECT PROPOSOL

on

FISH WEIGHT PREDICTION

Submitted in partial fulfilment of the requirement for the award of Degree of

Bachelor of Engineering

in

Computer Science and Engineering

Submitted by:

JOHNSON KUMAR PATEL	1NT19CS090
SOYUZ SHRESTHA	1NT19CS185
ABHASH KHANAL	1NT19CS007
AVAY KUSHWAHA	1NT19CS042

INTRODUCTION

Data mining is the process of finding anomalies, patterns, and correlations within large data sets to predict outcomes. Data mining techniques are deployed to scour large databases to find novel and useful patterns that might otherwise remain unknown. They also provide capabilities to predict the outcome of a future observation.

In fish market, it is important to determine the price of the fish correctly, and 2 important factors that impact on the price are fish species and the weight. Hence, the **objectives** of this projects are to predict the fish's weight based on the length, height, and width of the fish and to classify fish's species based on the weight, length, width, and height. The dataset chosen is Fish Market dataset was updated in 2019. The purpose of the dataset is to predict the fish species for the fish that caught off the coast of Finland.

The data mining task that we will be performing is regression using python in Anaconda. **Regression** is a data mining technique used to predict a range of numeric values (also called continuous values), given a particular dataset.

Data Mining Tasks

Regression:

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.

With Regression Method, we will use fish dataset to identify the relationship between weight with other numerical variables. We also try to see whether the weight of the fish can be predicted based on historical data.

DATA SET

This dataset is a record of 7 common different fish species in fish market sales. With this dataset, a predictive model can be performed using machine friendly data.

Let's look at the first rows of the dataset (Example)

	Species	Weight	Length1	Length2	Length3	Height	Width
0	Bream	242.0	23.2	25.4	30.0	11.5200	4.0200
1	Bream	290.0	24.0	26.3	31.2	12.4800	4.3056
2	Bream	340.0	23.9	26.5	31.1	12.3778	4.6961

3	Bream	363.0	26.3	29.0	33.5	12.7300	4.4555
4	Bream	430.0	26.5	29.0	34.0	12.4440	5.1340

Origin of data set: https://www.kaggle.com/aungpyaeap/fish-market.

Attributes:

Species : Nominal, Qualitative, Discrete

Weight : Interval, Quantitative, Continuous

Length : Interval, Quantitative, Continuous

Width : Interval, Quantitative, Continuous

Challenges:

Data access coupling:

Data access when tied to data store version and compute cause unique problems. The coupling results in issues when data store versions are updated or when compute processes change.

Methods and Model

Regarding the model we are trying to find out the best model by comparing their output and their performance in both the train and test instances. We will use various variations of regression model and a decision tree classifier to find out the optimal algorithm. Regression analysis allows us to quantify the relationship between outcome and associated variables. Many techniques for performing statistical predictions have been developed, but, in this project, two models-Multiple Linear Regression (MLR) and Decision tree regression are to be tested and compared.

ASSESSMENT

Before the phase of model creation, we will split our dataset into Train and Test dataset for more accurate model. We will use Train dataset to train our model, while the Test dataset will be used as a comparison whether our model can predict new data that has not been use. This will determine the validation of pattern discovered.

PRESENTATION AND VISUALIZATION

We will be using Microsoft PowerPoint to present our presentation via slides.

We will be using Anaconda graph to visualize the pattern discovered.

Roles

Mr. Johnson Kumar Patel

Role: Fish expert/ Coding

Mr. Soyuz Shrestha

Role: Coding Part

Mr. Avay Kushwaha

Role: Presentation / Report Creator

Mr. Abhash Khanal

Role: Data Mining / Proposal Creator

Schedule

TASKS	DATE
Proposal submission	29th DEC,2021
Report submission	17th JAN,2022
Presentation	17th JAN,2022

Bibliography

Data Set:

https://www.kaggle.com/aungpyaeap/fish-market

Introduction:

www.lifewire.com

Challenges:

https://medium.com/acing-ai/what-are-common-dataset-challenges-at-scale-6c440440d41d

Template:

https://github.com/vanivasudevan/Data-Mining/find/main

Other References:

https://hnslmp.medium.com/linear-regression-on-fish-market-dataset-using-python-eb2fc5f56aeb

https://www.geeksforgeeks.org/data-mining-data-attributes-and-quality/