 **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 LA2 REPORT**

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:

Soyuz Shrestha 1NT19CS185

Johnson Kumar Patel 1NT19CS090

Avay Kushwaha 1NT19CS042

Abhash Khanal 1NT19CS007

**Department of Computer Science and Engineering**

### ( NBA Accredited till 2020)

2020-2021

 **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

# DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

**CERTIFICATE**

# This is to certify that the Course Project titled “ *Fish Weight Prediction*” is an authentic work carried out by Soyuz Shrestha(1NT19CS185),Johnson Kumar Patel (1NT19CS090),Abhash Khanal(1NT19CS007), Avay Kushwaha(1NT19CS185) bonafide students of Nitte Meenakshi Institute of Technology, Bangalore in partial fulfilment for the award of the degree of *Bachelor of Engineering* in COMPUTER SCIENCE AND ENGINEERING of Visvesvaraya Technological University, Belagavi during the academic year *2019-2020.*

**Signature of the Faculty Incharge Name and Signature of the**

**HOD**

**(Dr. Vani V)**

**Dr. Saroja Devi**

**DECLARATION**

We hereby declare that

(i)This presentation does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged and the source being detailed in the report and in the reference’s sections.

(ii)All corrections and suggestions indicated during the interval presentation have been incorporated in the report.

(iii)Content of the report has been checked for the plagiarism requirement .

## ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crowned our effort with success. We express our sincere gratitude to our Principal **Dr. H. C. Nagaraj**, Nitte Meenakshi Institute of Technology for providing facilities.

We wish to thank our HoD**, Dr. Saroja Devi** for the excellent environment created to further educational growth in our college. We also thank him for the invaluable guidance provided which has helped in the creation of a better technical report.

Thanks to our Subject Faculty **Dr. Vani V** for guiding and helping us out in the successful completion of this presentation.

Signature

|  |  |
| --- | --- |
| Soyuz Shrestha | 1NT19CS185 |
| Johnson Kumar Patel | 1NT19CS090 |
| Abhash Khanal | 1NT19CS007 |
| Avay Kushwaha | 1NT19CS042 |

Date : 2022/01/17

**ABSTRACT**

About the project It is a supervised machine learning project built to predict the weight of fish based on various factors. We are using various data mining tasks to build a model that predicts the weight based on the factors line weight, height, length etc.

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.

**TABLE OF CONTENTS**

**Page Nos.**

**CHAPTER 1: INTRODUCTION-------------------------------------------------x**

* 1. Motivation---------------------------------------------------------------------------------------x
  2. Problem Domain--------------------------------------------------------------------------------x
  3. Aim and Objectives-------------------------------------------------------------------------------- x

**CHAPTER 2: Data Source and Data Quality---------------------------------x**

* 1. Dataset Used--------------------------------------------------x
  2. Data Pre-processing-----------------------------------------------------------------x

**CHAPTER 3: Methods and Models----------------------------x**

* 1. Data Mining Questions-----------------------------------------------------------------x
  2. Data Mining Algorithms----------------------------------------------------------------x
  3. Data Mining Models---------------------------------------------------------------------x

**CHAPTER 4: Model Evaluation & Discussion (with necessary visualization)**

**CHAPTER 5: Conclusion & Future Direction**

**CHAPTER 7: REFERENCES**

1. Introduction
   1. Motivation

Automated monitoring the growth of the fish is more convenient way of measuring and weighing it. 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 on historical data.

* 1. Problem Domain

Machine Learning (ML) which is subfield of data mining handles large scale well-formatted dataset efficiently. In the marketing field, machine learning can be used for estimate, measurement and prediction of various item. The main goal of this paper is to provide a tool for farmers to predict the weight of fish. This in turn will help to provide effective growth to fish. ML plays a very important role to detect the hidden discrete patterns and thereby analyse the given data. After analysis of data ML techniques help in fish weight prediction and measurement of fish which helps in this growth.

* 1. Aims and Objectives

The aim of this project is to develop a fish weight prediction system for intensive aquaculture setup that automatically predict the fish weight without manually weighing it. we develop a python algorithm that predicts the weights of the fish using some dataset of fish. This makes the monitoring of fish weight automated, reduce the time consumed by farmers and less stressful to fishes and helps to improve fish growth rate.

1. Data Source and Data Quality

2.1: Data Set used

The 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.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **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** |

2.1: Data Pre-processing

Regression:

Regression analysis is primarily used for two conceptually distinct purposes. First, regression analysis is widely used for [prediction](https://en.wikipedia.org/wiki/Prediction) and [forecasting](https://en.wikipedia.org/wiki/Forecasting), where its use has substantial overlap with the field of [machine learning](https://en.wikipedia.org/wiki/Machine_learning). Second, in some situation’s regression analysis can be used to infer [causal relationships](https://en.wikipedia.org/wiki/Causality) between the independent and dependent variables. Importantly, regressions by themselves only reveal relationships between a dependent variable and a collection of independent variables in a fixed dataset.

Multiple linear regression:

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. Multiple regression is an extension of linear (OLS) regression that uses just one explanatory variable.

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, dark, screen, screenshot

Description automatically generated

The dataset has been spitted into two parts the independent variable i.e. x and the dependent variable i.e. y and been fed into the training model.

1. **Methods and Models**

3.1: Data Mining Questions

* How does the data look like?

data**.**head()

Table

Description automatically generated

* Does the data have missing values?

data**.**isna()**.**sum()

Table

Description automatically generated

As you can see, in this case, since the problem is for beginners we do not have missing values, however, this is not always the case. In the following articles, we will have data with missing values and have the possibility to understand how to cope with that.

* What is the distribution of the numerical features?

data\_num **=** data**.**drop(columns**=**["Species"])  
  
fig, axes **=** plt**.**subplots(len(data\_num**.**columns)**//**3, 3, figsize**=**(15, 6))  
i **=** 0  
**for** triaxis **in** axes:  
 **for** axis **in** triaxis:  
 data\_num**.**hist(column **=** data\_num**.**columns[i], ax**=**axis)  
 i **=** i**+**1

Chart, histogram

Description automatically generated

As you can wee the distributions are okay. Target Variable(Weight) seems a little bit unbalanced and can be done some methods to balance the values.

* What is a correlation between the target variable and features?

plt**.**figure(figsize**=**(7,6))  
corr **=** data\_num**.**corr()  
sns**.**heatmap(corr,   
 xticklabels**=**corr**.**columns**.**values,  
 yticklabels**=**corr**.**columns**.**values, annot**=True**)  
plt**.**show()

Chart

Description automatically generated

The pairwise correlation of columns shows that all the numerical features have a positive correlation to the Weights. This means that the higher the lengths or width, the higher the weight is. It seems logical as well.

3.2: Data Mining Algorithms

* Supervised Machine Learning Algorithm
  1. Data Mining Models
* Multi-Linear Regression Model

**4: Model Evaluation & Discussion:**

**Graphical user interface, text, application, email

Description automatically generated**

The result of this analysis indicates that the multi-linear regression algorithm is an efficient algorithm with an r2 score of 0.93 for prediction of fish weight.

**Chart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

As we can see the model is a good fit for the selected dataset but there are a few outliers present in the prediction model.

**Conclusion:**

We develop an automated setup which can predict weight that makes use of predictive analysis which yields to a higher growth and survival rate. By avoiding stressors to fishes, such as capturing and handling to manually measure its weight and monitor its growth, the data gathered shows beneficent results to the rate of fish growth. The growth rate of the fishes in terms of weight improves. For future work, this study can be applied for growth monitoring of other animals on various environment.

For future work, this study can also be applied in predicting the weight and length of other aquatic animals in different aquaculture systems to determine its growth. This lessens the effort of the fish farmers and minimizes the possibilities of the fish to have a lower quality. Moreover, the acquired length and weight measurements may also be used to determine the number of fishes to cultivate in a given area to prevent space and food deprivation.

**Reflection Portfolio**

Looking back at the work we had to do to get this project done we can definitely say that we have learnt about various data mining algorithm as well as the models that are applicable for a certain type of data. We went from zero knowledge in implementing these algorithms to somewhat fulfilling the task at hand and we can say that this has greatly increase our understanding of the Data Mining topics.

Here are the things that we learnt during the preparation of this project as this report. We have learnt that we can do anything that we set out to do with a full heart and stop from doubting ourselves. This report shows that all we’ve researched and how we implemented the algorithms and compared their accuracies.

**Appendices**

1. Link to the dataset chosen

[Fish Market Dataset](https://www.kaggle.com/aungpyaeap/fish-market)

1. Python Codes Implemented

The Jupyter notebook file is attached.