**Data science programming**

**Data set of abalone**

**Name of team: -**

* **George salah adeb 41910271**
* **Fady magdy 42010349**
* **Asmaa hussein 42010749**
* **Hamza Mohamed khamis 41910094**
* **Abdelrahman Mamdouh 42010097**
* **Ahmed tarek gooda 41810257**

**Description: -**

**Relevant Information Paragraph: Predicting the age of abalone from physical measurements. The age of**

**abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task. Other measurements,**

**which are easier to obtain, are. used to predict the age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem.**

**Number of Instances:3983**

**Number of Attributes:9**

**Given is the attribute name, attribute type, the measurement unit and a**

**brief description. The number of rings is the value to predict either.**

**as a continuous value or as a classification problem.**

**Name Data Type Meas. Description**

**---- --------- ----- -----------**

**Sex nominal M, F, and I (infant)**

**Length continuous mm longest shell measurement**

**Diameter continuous mm perpendicular to length**

**Height continuous mm with meat in shell**

**Whole weight continuous grams whole abalone**

**Shucked weight continuous grams weight of meat**

**Viscera weight continuous grams gut weight (after bleeding)**

**Shell weight continuous grams after being dried.**

**Rings integer +1.5 gives the age in years.**

**Missing Attribute Values: not null**

**Problem Definition: -**

**Predicting the age of abalone from physical measurements. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem.**

**From the original data examples with missing values were removed (the majority having the predicted value missing), and the ranges of the continuous values have been scaled for use with an ANN (by dividing by 200).**

**Method: -**

**Data collection: The data set used for this analysis is the “abalone” data set obtained from Kaggle. Contains information on the age of the abalone which is determined by cutting the shell through the cone.**

**Data Cleaning: The dataset has been preprocessed to handle missing values, remove irrelevant columns, remove duplicated data, Handling Outliers and ensure data quality and consistency.**

**Exploratory Data Analysis (EDA): An exploratory analysis was performed to obtain a count of the number of rings through a microscope - a tedious and time-consuming task. Other measurements, which are easier to obtain, are used to predict age.**

**Feature Selection: Based on EDA, the specific features for prediction were budget and popularity of films. These features are the most relevant in determining a movie's revenue.**

**Model Selection: Linear regression was chosen as the predictive modeling technique because of its simplicity and its ability to capture linear relationships between variables. The scikit-learn library in Python was used to implement the linear regression model.**

**Model training: The dataset was divided into training and test sets. The training set was used to fit the linear regression model to the data.**

**Model Evaluation: The trained model was evaluated using the Mean Squared Error (MSE) measure, which measures the mean squared difference between expected and actual revenue values.**

# Experiment: -

# Data cleaning: Missing values were handled by insertion or deletion. Discordant rows were removed, and data quality checks were performed to ensure accuracy.

# Exploratory Data Analysis: Graphics such as scatterplots, histograms, and correlation matrices were created to explore relationships among variables and understand the distribution of data.

# Training and performance evaluation: A linear regression model was trained on the training set and evaluated using the MSE metric. The performance of the model was evaluated by comparing the predicted revenue with the actual revenue values in the pilot test.

# References: -

# Kaggle: ear data. removed from https://www.kaggle.com/datasets/ursmaheshj/top-10000-popular-movies-tmdb-05-2023.

# This report presents an analysis of the estimated crustal age determined by cutting the crust through the cone. using linear regression. The findings and findings can be used by industry professionals, analysts, or researchers to understand which factors use other easily obtained measurements to predict age.