UNIVERSITI TEKNOLOGI MALAYSIA

SECP2753

DATA MINING

ASSIGNMENT 1

Title: Data Exploration and Preprocessing of CPU Benchmark and Abalone Datasets

PROF. MADYA DR. ROLIANA BINTI IBRAHIM

NO	NAME	MATRIC NO
1.	CHUA JIA LIN	A23CS0069
2.	EVELYN GOH YUAN QI	A23CS0222

1. Introduction

Data mining is a strong technique for identifying relevant patterns and insights in massive datasets from various domains.This report focuses on two datasets which are the Computer Hardware dataset, which includes specifications and performance scores of computer processors, and the Abalone dataset, which contains physical attributes of abalones for age prediction.

The aim of this report is to analyze both datasets to discover relevant insights and patterns, followed by applying essential preprocessing techniques, specifically normalization and discretization, to one selected dataset in preparation for data mining model development.

By combining exploratory research and data preparation, this report emphasizes the significance of understanding and manipulating data before applying predictive techniques.

2. Datasets Insight

Abalone Dataset

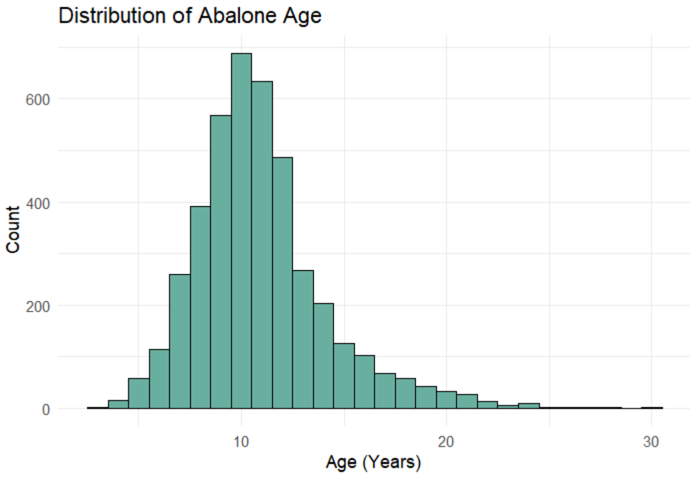
The abalone dataset contains 4177 records, each representing physical measurements of individual abalones. The dataset consists of eight features including length, diameter, height, weight, and one target attribute which is the number of rings on the shell. The primary purpose of this dataset is to predict the age of abalone from physical measurements collected from the abalones. The dataset's attributes includes:

- **Sex** (Categorical)
  Gender of the abalone: M (Male), F (Female), I (Infant)
- **Length** (Numerical)
  Longest shell measurement (in mm)
- **Diameter** (Numerical)
  Perpendicular to length (in mm)
- **Height** (Numerical)
  With meat in shell (in mm)
- **Whole Weight** (Numerical)
  Weight of whole abalone (in grams)
- **Shucked Weight** (Numerical)
  Weight of meat (in grams)
- **Viscera Weight** (Numerical)
  Gut weight (in grams)
- **Shell Weight** (Numerical)
  After being dried (in grams)

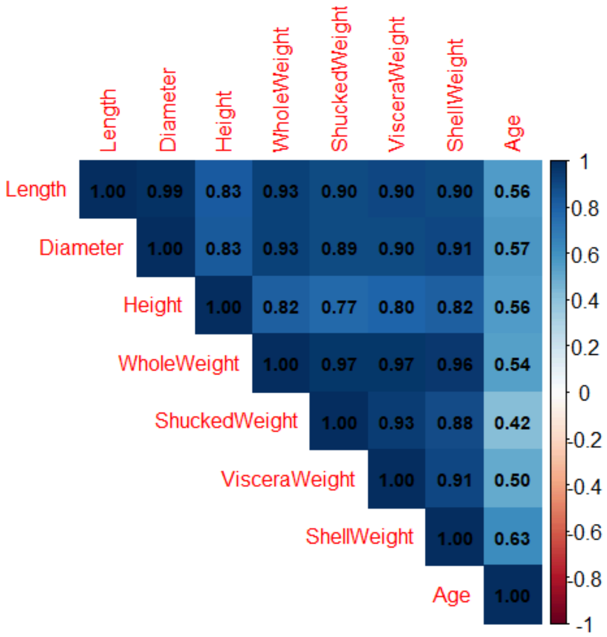
- **Rings** (Numerical)
  Number of rings on the shell (+1.5 gives the age in years)

Key Insights:

1. **Distribution of Abalone Age**
   The number of rings acts as a proxy for the age of an abalone using the formula Age = Rings + 1.5. Most abalones fall between 6 and 15 years old. The histogram is skewed to the right, indicating that there are many young abalones and fewer older abalones.

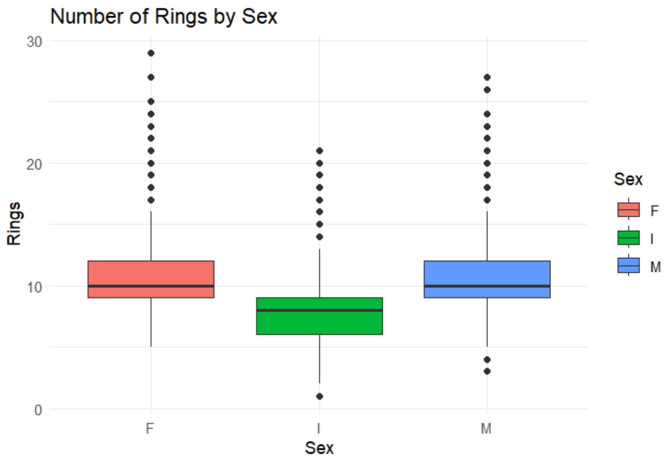


2. **Feature Correlation**
   The size and weight features like length, diameter, and shell weight are strongly related. The bigger the size of abalones, the greater the weight of abalones. Shell weight, whole weight, and length are also helpful in predicting abalone's age.



3. **Gender vs Rings**

   Gender does not strongly affect the age of abalone. However, the box plot shows that infant abalones are obviously younger than male and female abalones. Males and females have similar sizes and ages.



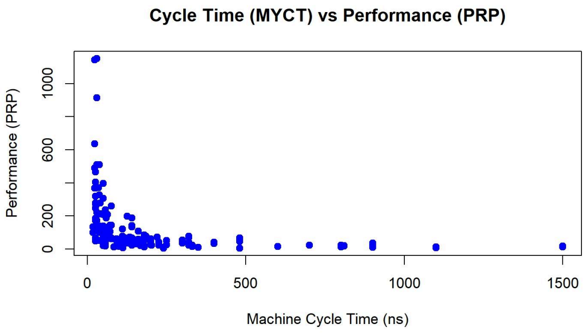
<u>Computer Hardware Dataset</u>

The computer hardware dataset contains 209 entries including information about various computer hardware configurations, with a primary focus on CPU related specifications and their performance ratings. The dataset is intended to assist people learn how different hardware components affect computer system performance. The dataset includes processor-related attributes such as:

- **Vendor Name** (Categorical)
  - Indicates the manufacturer of the CPU (e.g., adviser, amdahl, cray).
- **Model Name** (Categorical)
  - Specifies the CPU model identifier.
- **MYCT (Machine Cycle Time)** (Numerical)
  - Represents the time required to complete one CPU cycle.
- **MMIN (Minimum Main Memory)** (Numerical)
  - Minimum amount of main memory required by the system.
- **MMAX (Maximum Main Memory)** (Numerical)
  - Maximum supported main memory for the CPU.
- **CACH (Cache Memory Size)** (Numerical)
  - Indicates the size of the CPU's cache memory.
- **CHMIN (Minimum Channels)** (Numerical)
  - Minimum number of data channels supported by the CPU.
- **CHMAX (Maximum Channels)** (Numerical)
  - Maximum number of data channels supported.
- **PRP (Published Relative Performance)** (Numerical)
  - Target variable representing the overall performance score of the CPU.

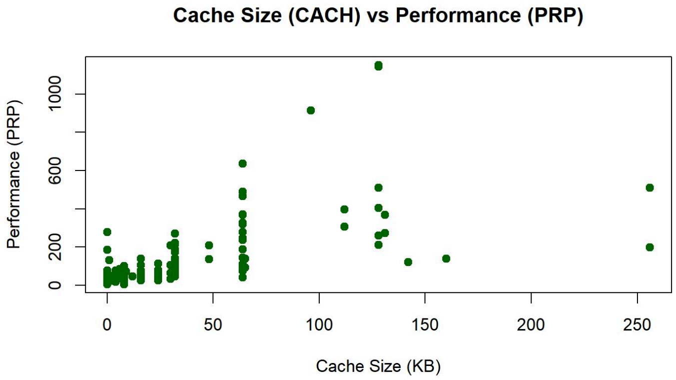
Key Insights:

1. **Cycle Time (MYCT) vs Performance (PRP)**
   The scatter plot of machine cycle time (MYCT) against performance (PRP) reveals a negative trend, where CPUs with lower cycle times generally achieve higher performance scores. This supports the idea that faster processors execute instructions more efficiently. However, the relationship is not perfectly linear—some CPUs with moderate MYCT still outperform faster ones, indicating that other factors such as cache size and memory configuration also influence overall performance.



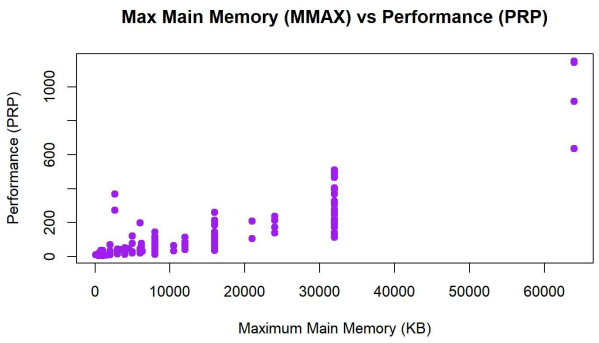
2. **Cache Size (CACH) vs Performance (PRP)**

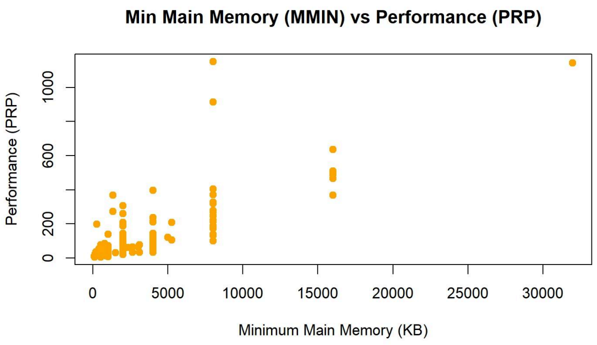
   The plot of cache size versus performance shows a moderate positive relationship, suggesting that CPUs with larger cache memory tend to perform better. Since the cache stores frequently accessed data close to the processor, it significantly reduces memory access time. This relationship validates cache size as a valuable predictor in performance modeling, although the effect may vary depending on other hardware components.



3. Main Memory Size (MMIN/MMAX) vs Performance

Scatter plots for both MMIN (minimum memory) and MMAX (maximum memory) show strong positive correlations with PRP. This indicates that CPUs equipped with larger memory capacities consistently deliver better performance. The data points demonstrate a clear upward trend, reinforcing the importance of RAM size in boosting CPU efficiency. These insights confirm that both minimum and maximum supported memory are critical factors in system performance evaluation.





3. Preprocessing Dataset : Abalone

Among the abalone dataset and computer hardware dataset, we decided to choose the abalone dataset as preparation for our data mining model development. The dataset contains various continuous numerical features, making it suitable for preprocessing methods such as normalization and discretization. These techniques can ease the data processing and improve performance of data mining tools.

4. Preprocessing Techniques

1. Normalization

Normalization is a preprocessing technique used to scale numerical data into a standard range, typically [0, 1]. This is important for ensuring that attributes with larger numeric ranges do not dominate those with smaller ranges in data mining models.

In this report, we use **Min-Max Normalization**, which maps an original value v to a new value v′ using the formula:

$$v' = \frac{v - \min_A}{\max_A - \min_A} \times (\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}_A$$

Example 1: If the minimum and maximum values for ***Length*** are 0.075 and 0.815 respectively, and one abalone has a length of 0.455

v′ = ((0.455- 0.075) ÷ (0.815-0.075)) x (1 - 0) + 0

= 0.5135

Thus, the normalized Length value is approximately **0.5135**.

Example 2: If the minimum and maximum values for ***Whole Weight*** are 0.002 and 2.8255 respectively, and one abalone has a length of 0.798

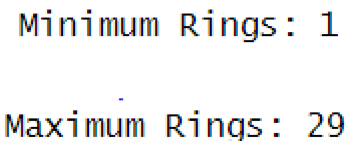
v′ = ((0.798 - 0.002) ÷ (2.8255 - 0.002)) x (1 - 0) + 0

= 0.2819

Thus, the normalized value of Whole Weight is **0.2819**.

2. Discretization

Discretization is a technique that converts continuous data such as age or weight into categories or bins. This technique is useful for simplifying data interpretation. For the abalone dataset, we decided to discretize the Rings of abalones using equal-width binning.



```
Minimum Rings: 1

Maximum Rings: 29
```

Let:

Lowest value(A) = 1

Highest value (B) = 29

No. of Bins (N) = 3

Bin Width = (B - A) / N

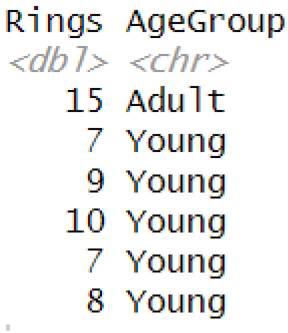
= (29 - 1) / 3

= 28 / 3

= 9

Rings of abalone (+1.5 gives the age in years) are divided into three age groups which are:

- **Young**: [1 , 10] include Rings >= 1 and Rings <= 10
- **Adult**: [10 , 19] include Rings > 10 and Rings <= 19
- **Old**: [19 , 29] include Rings > 19 and Rings <= 29



```
Rings AgeGroup
<dbl> <chr>
   15 Adult
    7 Young
    9 Young
   10 Young
    7 Young
    8 Young
```

Example 1: An abalone with **5 rings** and age of 6.5 falls into the **Young** category.

Example 2: An abalone with **11 rings** and age of 12.5 falls into the **Adult** category.

Example 3: An abalone with **20 rings** and age of 21.5 falls into the **Old** category.

5. Discussion

The preprocessing process plays a vital role in preparing raw data for effective data mining model development. In this assignment, we applied two key preprocessing techniques which are normalization and discretization to the Abalone dataset, with each contributing uniquely to data quality enhancement.

Min-max normalization was applied to the continuous numerical features such as length and whole weight to scale them within a range of [0,1]. This ensures uniformity in feature scales, which is especially important for distance-based algorithms like k-nearest neighbors and clustering approaches. Without normalization, features with greater numeric ranges may outperform others, resulting in biased or inferior model results. The modified values preserve the distribution shape while normalizing the range, making the data suitable for mining tasks.

Discretization using equal-width binning was applied to the Rings feature of the Abalone dataset. For the Rings attribute, we grouped the integer values into three categories: Young (Rings <= 10), Adult (Rings between 10 and 19), and Old (Rings > 19). This makes it easier to analyze patterns across life stages and supports classification tasks.

By combining both techniques, we improved data uniformity, interpretability, and suitability for data mining tasks. Additionally, by reducing potential noise and outlier effects, these preprocessing steps improve the accuracy and reliability of the model.

6. Conclusion

In this assignment, we explored two datasets, which are the Abalone dataset and the Computer Hardware dataset to learn the insights of the datasets and prepare the data for mining tasks. After analysing the datasets, we discovered important patterns of the datasets, such as the effect of physical features towards abalone's age and the impact of hardware specifications on CPU performance.

After analysing the two datasets, we chose the Abalone dataset for further mining tasks. We applied normalization and discretization on the dataset. Normalization helps to bring all numerical values to a smaller scale, making the data more suitable for machine learning algorithms, whereas discretization helps to group continuous values into meaningful categories to simplify the analysis process and enable classification modelling.

In conclusion, the exploration and preprocessing steps help to improve the quality and usability of the data. These steps are essential in the data mining process, as they can discover hidden patterns, improve the model accuracy, and lead to more reliable decision making. This assignment highlights the importance of preparing data properly before applying other advanced analytic techniques on the data.

References

1. *UCI Machine Learning Repository*. (n.d.). Archive.ics.uci.edu. https://archive.ics.uci.edu/dataset/1/abalone
2. Meek, C. (n.d.). *Computer Hardware Data Set*. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/29/computer+hardware
3. Han, J., & Kamber, M. (2012). *Data mining : concepts and techniques*. Elsevier/Morgan Kaufmann.
4. Hu, Y. (2024, September 11). *Abalone Age Prediction and Classification with labels*. Medium. https://medium.com/@hy1550278246/abalone-age-prediction-and-classification-with-labels-2ec612e26d4f