

Data and Data Preprocessing

Problem 1: Types of Attributes (14 points)

Classify the following attributes as nominal, ordinal, interval, ratio. **Explain why.**

(a) Rating of an Amazon product by a person on a scale of 1 to 5

- Rating of the product follow distinctness property.
- The scale has an inherent ordering, i.e., a rating of 5 is higher than a rating of 4, and a rating of 4 is higher than a rating of 3 and so on.
- Addition doesn't show the same preference between intervals, i.e., (a rating of 3 - a rating of 2) does not equal (a rating of 5 - a rating of 4).
- Therefore, this is an **ordinal** attribute.

(b) The Internet Speed

- Internet speed is follow distinctness property.
- Speed has an inherent ordering, i.e., 3 Mbps is faster than 2 Mbps, and 2Mbps is faster than 1 Mbps.
- Addition shows the same preference between intervals, i.e., a difference of 1 Mbps means equal change in speed.
- Terms like "twice as fast" are meaningful. For example, a speed of 2 Mbps is twice the speed of 1 Mbps.
- Therefore, this is a **ratio** attribute.

(c) Number of customers in a store.

- Number of customers are follow distinctness property.
- Number of customers has an inherent ordering, i.e., 30 customers is more than 20 customers.
- Addition shows the same preference as other intervals, i.e., 20 customers + 10 customers = 25 customers + 5 customers.
- Terms like "twice, half" make sense. For example, 10 customers is half of 20 customers.
- Therefore, this is a **ratio** attribute.

(d) UCF Student ID

- Student ID follow distinctness property.
- There is no unique ordering between student ID's
- Therefore, this is a **nominal** attribute.

(e) Distance

- Distance follows distinctness property.
- Distance has an inherent ordering, i.e., 30 miles is more than 20 miles.
- Addition shows the same preference as other intervals, i.e., 20 miles + 10 miles = 25 miles + 5 miles.
- Terms like "twice, half" make sense. For example, 10 miles is half of 20 miles.
- Therefore, this is a **ratio** attribute.

(f) Letter grade (A, B, C, D)

- Letter grade follows distinctness property.
- Letter grade has an inherent ordering, i.e., A grade is higher than B grade.
- Addition doesn't show the same preference between intervals, i.e., (B grade - C grade) does not necessarily equal (A grade - B grade).
- Therefore, this is an **ordinal** attribute.

(g) The temperature at Orlando

- Temperature follows distinctness property.
- Temperature has an inherent ordering, i.e., 80 Fahrenheit is hotter than 70 Fahrenheit.
- Addition shows the same preference across intervals, i.e., (70 Fahrenheit + 10 Fahrenheit) is the same as 80 Fahrenheit.
- Terms like "twice as hotter" doesn't make sense. For example, 80 Fahrenheit is not twice as hot as 40 Fahrenheit.
- Therefore, this is an **interval** attribute.

Problem 2: Exploring Data Preprocessing Techniques (26 points)

Read the solution post of the Kaggle Titanic Dataset:

<https://www.kaggle.com/code/preejababu/titanic-data-science-solutions>. Run the code and reproduce the data preprocessing and classification modeling steps.

Q1 (Reproduce): Please read, understand, run the code and reproduce the model accuracies. Please briefly explain whether you can reproduce the classification accuracies of 'Support Vector Machines', 'KNN', 'Logistic Regression', 'Random Forest', 'Naive Bayes', 'Perceptron', 'Stochastic Gradient Decent', 'Linear SVC', 'Decision Tree'. (10 points)

In the given Kaggle Titanic dataset, the main aim of the dataset is to find the survival rate. The workflow stages used to process the data using models are Classifying, Correlating, Converting, Completing, Correcting, Creating and Charting. Initially, the dataset contains the following features: 'PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Cabin' and 'Embarked'. After running a few scenarios, features such as 'Ticket', 'Cabin' and 'PassengerId' were dropped without creating any new features. Some features such as 'Name', 'Age', 'Parch', 'SibSp' and 'Fare' were modified to create new features such as 'Title', 'AgeBand', 'FamilySize', 'IsAlone' and 'FareBand'.

I tried reproducing the code with the same ML models and I was able to see the same accuracies for all the algorithms for every run except for the Stochastic Gradient Descent. Since the Stochastic Gradient Decent is an iterative algorithm which takes the datasets randomly for each run, this algorithm will display a different score for each run.

ALGORITHM	ACCURACY (RUN 1)	RANK (RUN 1)	ACCURACY (RUN 2)	RANK (RUN 2)
Random Forest	86.76	1	86.76	1
Decision Tree	86.76	2	86.76	2
KNN	84.74	3	84.74	3
Support Vector Machines	83.84	4	83.84	4
Logistic Regression	80.36	5	80.36	5
Linear SVC	79.12	6	79.12	6
Perceptron	78.00	7	78.00	7
Stochastic Gradient Descent	73.63	8	71.72	9
Naive Bayes	72.28	9	72.28	8

Q2 (Improve): Is the data preprocessing process proposed in the Kaggle post the best preprocessing solution? If yes, please explain why. If not, can you leverage what you learned in the class and your previous experiences to improve data processing, to obtain better accuracies for all these classification models? Describe what is your improved data preprocessing, and what are your improved accuracies? (16 points)

No, I think that the data preprocessing process proposed in the Kaggle post is not the best preprocessing solution. The plot between 'Age' and 'Survived' reveals that many passengers in the age group between 20 and 30 did not survive. Similarly, the plot between 'Fare' and 'Survived' shows that many passengers who paid a fare of less than 10.5 did not survive and that

many passengers who paid a fare between 10.5 and 15 survived. In order to accommodate this observation, I increased the number of buckets from 5 to 7 for the 'AgeBand' feature and increased the number of buckets from 4 to 5 for the 'FareBand' feature.

The changes I made to the solution of the Kaggle post increased the accuracy from 86.76 to 88.33 for the best performing algorithms. Previously, higher difference to the age values was given to AgeBand and FareBand. I decreased the interval length and added new values to enable the code to run faster with high accuracy. Below is the accuracy of the algorithms after making changes to the code.

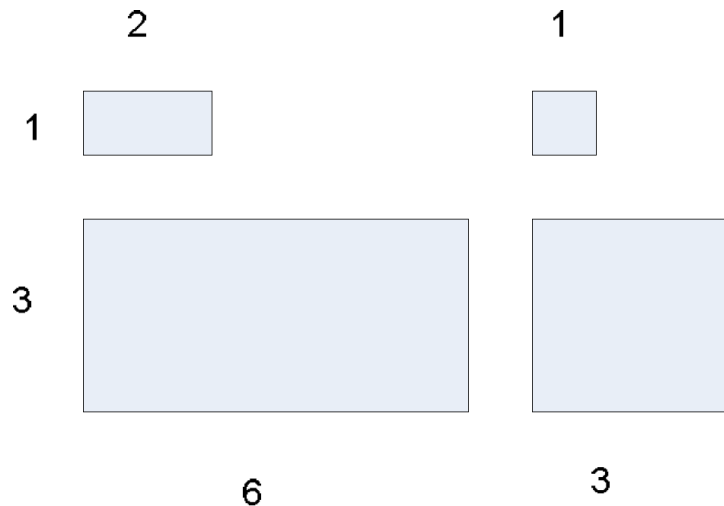
ALGORITHM	ACCURACY (KAGGLE)	RANK (KAGGLE)	ACCURACY (NEW)	RANK (NEW)
Random Forest	86.76	1	88.33	1
Decision Tree	86.76	2	88.33	2
KNN	84.74	3	86.08	3
Support Vector Machines	83.84	4	84.51	4
Logistic Regression	80.36	5	79.35	6
Linear SVC	79.12	6	78.68	7
Perceptron	78.00	7	79.80	5
Stochastic Gradient Descent	73.63	8	73.96	9
Naive Bayes	72.28	9	73.96	8

The below link contains the code with the changes made to AgeBand and FareBand features that resulted in the above improved accuracies.

<https://github.com/Viknesh-Rajaramon/Data-Mining/blob/master/HW1/HW1.ipynb>

Problem 3: Distance/Similarity Measures (10 points)

Given the four boxes shown in the following figure, answer the following questions. In the diagram, numbers indicate the lengths and widths and you can consider each box to be a vector of two real numbers, length and width. For example, the top left box would be (2,1), while the bottom right box would be (3,3). Restrict your choices of similarity/distance measure to Euclidean distance and correlation. **Please explain your choice.**



Which proximity measure would you use to group the boxes based on their shapes (length-width ratio)?

Let us consider the box sizes as their coordinates (x, y). So,
Box 1 = (2, 1), Box 2 = (1, 1), Box 3 = (6, 3), Box 4 = (3, 3)

We can use correlation measure for grouping the boxes by shape. We can see that between box 1 and box 3, as the length increases, the width also increases by the same, i.e., the length-width ratio remains the same for both the boxes. Between box 1 and box 2, though width is same, the length is different, i.e., the length-width ratio is different for both the boxes.

Which proximity measure would you use to group the boxes based on their size?

Let us consider the box sizes as their coordinates (x, y). So,
Box 1 = (2, 1), Box 2 = (1, 1), Box 3 = (6, 3), Box 4 = (3, 3)

$$\text{Euclidean distance } (d) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

We can use Euclidean distance measure for grouping the boxes by size since the variables are independent and do not have any correlation between them.

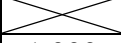
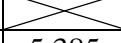
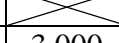

$$\begin{aligned}\text{Euclidean_Distance}(\text{Box 1, Box 2}) &= \sqrt{(1 - 2)^2 + (1 - 1)^2} = \sqrt{(-1)^2 + (0)^2} = \sqrt{1} = 1.000 \\ \text{Euclidean_Distance}(\text{Box 1, Box 3}) &= \sqrt{(6 - 2)^2 + (3 - 1)^2} = \sqrt{(4)^2 + (2)^2} = \sqrt{20} = 4.472\end{aligned}$$

$$\text{Euclidean_Distance}(\text{Box 1, Box 4}) = \sqrt{(3 - 2)^2 + (3 - 1)^2} = \sqrt{(1)^2 + (2)^2} = \sqrt{5} = 2.236$$

$$\text{Euclidean_Distance}(\text{Box 2, Box 3}) = \sqrt{(6 - 1)^2 + (3 - 1)^2} = \sqrt{(5)^2 + (2)^2} = \sqrt{29} = 5.385$$

$$\text{Euclidean_Distance}(\text{Box 2, Box 4}) = \sqrt{(3 - 1)^2 + (3 - 1)^2} = \sqrt{(2)^2 + (2)^2} = \sqrt{8} = 2.828$$

$$\text{Euclidean_Distance}(\text{Box 3, Box 4}) = \sqrt{(3 - 6)^2 + (3 - 3)^2} = \sqrt{(-3)^2 + (0)^2} = \sqrt{9} = 3.000$$

	Box 1	Box 2	Box 3	Box 4	Minimum Distance
Box 1		1.000	4.472	2.236	1.000
Box 2	1.000		5.385	2.828	1.000
Box 3	4.472	5.385		3.000	3.000
Box 4	2.236	2.828	3.000		2.236

Similar the size of the boxes, smaller is the Euclidean distance. From the above table, we can see that for Box 1, the Euclidean distance is minimum with Box 2. Similarly for Box 3, the Euclidean distance is minimum with Box 4. Therefore, Box 1 and Box 2 are similar, Box 3 and Box 4 are similar based on size.