

**CSEN911: Data Mining**  
Winter Term 2023  
**Problem Set 1 – Data Preprocessing**

**Problem 1**

Suppose a hospital tested the age and body fat data for 9 randomly selected adults with the following result:

|             |     |      |     |      |      |      |      |      |      |
|-------------|-----|------|-----|------|------|------|------|------|------|
| <i>age</i>  | 23  | 23   | 27  | 27   | 39   | 41   | 47   | 49   | 50   |
| <i>%fat</i> | 9.5 | 26.5 | 7.8 | 17.8 | 31.4 | 25.9 | 27.4 | 27.2 | 31.2 |

Given  $\sigma_{age} = 11.29$ , and  $\sigma_{\%fat} = 8.91$

- Calculate the mean of age and %fat.
- Normalize the two variables based on the z-score normalization. (  $new\ v_i = \frac{v_i - \bar{A}}{\sigma_A}$  )
- Mention whether these two attributes are positively or negatively correlated or not correlated at all, if the correlation coefficient is 0.71

## Problem 2

Jake did a survey of the numbers of brothers and sisters of the children in his class, and summarized the results in the following table:

| # Siblings | Frequency |
|------------|-----------|
| 0          | 91        |
| 1          | 93        |
| 2          | 93        |
| 3          | 98        |
| 4          | 75        |
| 5          | 79        |
| 6          | 90        |
| 7          | 85        |
| 8          | 50        |
| 9          | 5         |

Draw the boxplot for these observations and identify outliers if present.

## Problem Set 2 – Regression

### Problem 3

Most supermarkets use scanners at the checkout counters. The data collected this way can be used to evaluate the effect of price and store's promotional activities on the sales of any product. The promotions at a store change weekly, and are mainly of two types: flyers distributed outside the store and through newspapers (which may or may not include that particular product), and in-store displays at the end of an aisle that call the customers' attention to the product. Weekly data was collected on a particular product brand, including sales (in number of units), price (in dollars), flyer (1 if product appeared that week, 0 if it didn't) and display (1 if a special display of the product was used that week, 0 if it wasn't).

- a) As a preliminary analysis, a simple linear regression model was done. The fitted regression equation was:  $sales = 2259 - 1418 price$ .

- 1) The response variable is:

|              |                          |
|--------------|--------------------------|
| i. price.    | <input type="checkbox"/> |
| ii. flyer.   | <input type="checkbox"/> |
| iii. sales.  | <input type="checkbox"/> |
| iv. display. | <input type="checkbox"/> |

- 2) According to this model, how many units will be sold, on average, when the price of the product is \$1.10?

|              |                          |
|--------------|--------------------------|
| i. 3818.8.   | <input type="checkbox"/> |
| ii. 699.2.   | <input type="checkbox"/> |
| iii. 1066.9. | <input type="checkbox"/> |
| iv. 3902.9.  | <input type="checkbox"/> |

b) Next, the categorical variables flyer and display were added to the model. The fitted regression equation was:

$$sales = 3829.5 - 5056 price + 804.12 flyer - 31.49 display.$$

1) According to this model, what is the average effect of advertising the product on the weekly flyer, after adjusting for price and display?

|   |                          |
|---|--------------------------|
| i. There will be a significant increase in sales of about 804 units.      | <input type="checkbox"/> |
| ii. There will be an insignificant increase in sales of about 9 units.    | <input type="checkbox"/> |
| iii. There will be a significant decrease in sales of about 804 units.    | <input type="checkbox"/> |
| iv. There will be an insignificant decrease in sales of about 5056 units. | <input type="checkbox"/> |

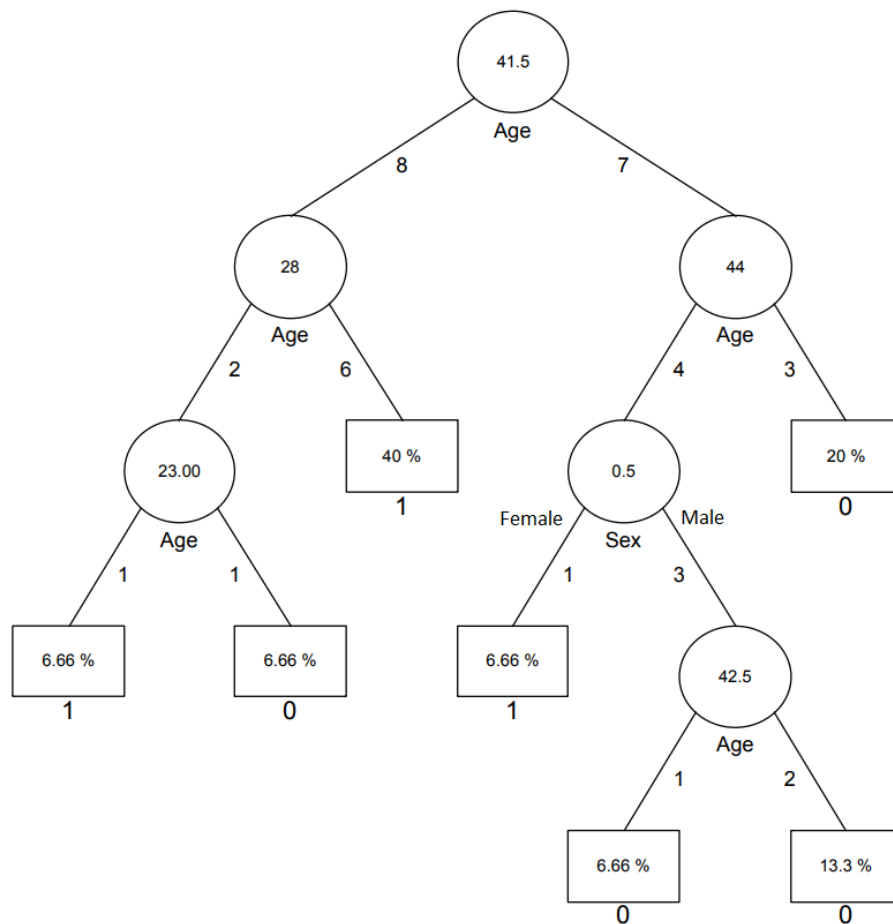
2) Flyers usually advertise products that are on sale that week. This implies that we should add to the model:

|  |                          |
|--|--------------------------|
| i. an interaction between display and flyer.   | <input type="checkbox"/> |
| ii. an interaction between price and flyer.    | <input type="checkbox"/> |
| iii. an interaction between price and display. | <input type="checkbox"/> |
| iv. an additional term for flyer.              | <input type="checkbox"/> |

## Problem Set 3 – Classification

### Problem 4

A credit card company has created a classification tree shown below for a promotion consisting of a special offer on credit card insurance. The class **1** corresponds to customers who responded, class **0** consists of those who did not. The company will use the tree to send promotion offers to dormant customers to persuade them to begin to use the card. Age, sex and income were used as input variables for these customers.



The company would like to have a few, simple, English language rules that embody the decisions represented by the tree. Write out succinctly the rules you would suggest.

## Problem 5

Use the Naive Bayes algorithm to classify the following customers into one of the two categories: Default (Yes) or No Default (No).

| Customer | Income | Employment Status | Loan Default |
|----------|--------|-------------------|--------------|
| 1        | Low    | Employed          | No           |
| 2        | Medium | Self-Employed     | No           |
| 3        | High   | Employed          | No           |
| 4        | Low    | Unemployed        | Yes          |
| 5        | Medium | Employed          | No           |
| 6        | Low    | Self-Employed     | Yes          |
| 7        | High   | Self-Employed     | No           |
| 8        | Medium | Unemployed        | Yes          |

**a) Customer 9: Income = Low, Employment Status = Employed**

**b) Customer 10: Income = Medium, Employment Status = Self-Employed**