Machine Learning Exercise - 1

Saurabh Jondhale - 3574931
st1802882@stud.uni-stuttgart.de

1 Task-1a

Handling Missing data			
ID	Age	Income(k)	Owns Cars
1	25	50	Yes
2	33.75	40	No
3	35	55	Yes
4	45	70	Yes
5	30	60	No

Handling Missing data			
ID	Number of Vehicles	Preferred transport mode	Income scaled
1	2	Car	0.333333
2	0	Public Transport	0.000000
3	1	Car	0.500000
4	0	Car	1.000000
5	0	Bike	0.666667

2 Task-1b

$scaled value = \frac{value - min}{max - min}$
$Using Min-Max scaling on the "Income(K)"$ column with the minimum income of 40K $and the maximum of 70K$, the scaled values are calculated as shown in the "Income (K)$Scaled"$ $column of the updated dataset above.$

3 Task-1c

Solution:
Binary Encoding for "Owns Car"

1. $Yes \rightarrow 0$
2. $No \rightarrow 1$

One-Hot Encoding for "Preferred Transport Mode" is as follows:

1. Car: $[1, 0, 0]^T$
2. Public Transport: $[0, 1, 0]^T$
3. Bike: $[0, 0, 1]^T$

The dataset would be extended with additional columns to accommodate the one-hot encoded "Preferred Transport Mode", resulting in a structure as described in the following table:

Owns Car (Encoded)	Car	Public Transport	Bike
1	1	0	0
0	0	1	0
1	1	0	0
1	1	0	0
0	0	0	1

4 Task-2a

1. Calculate the Euclidean distance using the equation:

$$\sqrt{(x_{1i} - x1j^2) - (x_{2i} - x2j^2 - (x_{3i} - x3j^2))} \tag{1}$$

Distance to Observation = $1.414, 2.236, 2.449, 2.000, 1.414$

2. Identify the three nearest neighbors based on the smallest distances.
- The three nearest neighbors are Observations $d(O_{1,4,5}) = 1.414, 2.000, 1.414$ respectively.

3. Determine the majority class among these neighbors.
- Observation 1's class: 0 - Observation 4's class: 1 - Observation 5's class: 1
- The majority class among the nearest neighbors is Class 1. Therefore, if we classify the new observation using the KNN algorithm with K=3, it would be assigned to Class

5 Task-2b

Solution: For this question, we analyze how the classification of the new observation (X1=3, X2=3, X3=2) changes with K=1 and K=5, and discuss the implications of choosing different K values. When K=1 1. Find the nearest single neighbor: From the previous calculations, we know that the nearest neighbors (closest distances) are Observations 1 and 5, both with distances of 1.414. For K=1, we can choose either of them (let's use Observation 1 for this example). 2. Class of the nearest neighbor: Observation 1 belongs to Class 0. Classification with K=1: The new observation would be classified as Class 0. When K=5 1. Find the five nearest neighbors: All observations in our dataset will be considered, as we only have five observations. 2. Determine the majority class among these neighbors: We have two observations belonging to Class 0 (Observations 1 and 2) and three belonging to Class 1 (Observations 3, 4, and

5) Classification with K=5: The new observation would be classified as Class 1. Benefits and Drawbacks: - Smaller K (K=1): More sensitive to noise in the dataset; the classification is based on the nearest observation only, which can lead to overfitting. - Larger K (K=5): Tends to smooth out the classification and reduce the effect of noise. However, it may include points that are farther away, which could lead to underfitting or misclassifying the new observation if the boundary between classes is complex.

6 Task-2c

Solution: With distance-weighted voting, the influence of each neighbor on the classification decision is weighted by the inverse of their distance to the point being classified, giving closer neighbors more influence. 1. Calculate weighted influence for the three nearest neighbors: - Weight for Observation 1: 0.707 - Weight for Observation 4: 0.5 - Weight for Observation 5: 0.707 2. Calculate weighted votes for each class: - Total weighted vote for Class 0: 0.707 (from Observation 1) - Total weighted vote for Class 1: 1.207 (from Observations 4 and 5) The new observation would be classified as Class 1, because the total weighted vote for Class 1 is higher than that for Class 0.