Maggie Lin

mclin

1 Evaluation Measures (11 points)

1 a)

BTF	Gender	EL	MS	Label	**Left**
Travel Rarely	Male	L2	Married	N	**No**
Travel Rarely	Male	L2	Married	N	**No**
Travel Rarely	Female	L4	Single	N	**No**
Travel Rarely	Male	L3	Divorced	Y	**No**
Travel Rarely	Female	L3	Single	N	**No**
Travel Frequently	Male	L1	Divorced	N	**Yes**
Travel Frequently	Male	L1	Single	N	**No**
Travel Frequently	Female	L2	Married	Y	**Yes**
Travel Frequently	Female	L4	Divorced	Y	**No**
Travel Frequently	Male	L1	Married	Y	**Yes**
Travel Frequently	Female	L2	Married	Y	**Yes**
Travel Frequently	Male	L2	Single	N	**No**
Travel Frequently	Female	L1	Single	Y	**Yes**
Travel Frequently	Male	L4	Single	Y	**No**
Travel Frequently	Female	L3	Divorced	N	**No**
Travel Frequently	Male	L4	Divorced	N	**Yes**

	Predicted Yes	**Predicted No**
Actual Yes	**4**	**3**
Actual No	**2**	**7**

1 b)

Accuracy:

$$\text{Accuracy: } \frac{TP+TN}{TP+TN+FP+FN} = \frac{4+7}{4+7+2+3} = \frac{11}{16} = 68.75\%$$

Error Rate:

$$\text{Error Rate: } \frac{FP+FN}{TP+TN+FP+FN} = \frac{2+3}{4+7+2+3} = \frac{5}{16} = 31.25\%$$

Precision:

$$\text{Precision: } \frac{TP}{TP+FP} = \frac{4}{4+2} = \frac{4}{6} = 66.67\%$$

Recall:

$$\text{Recall: } \frac{TP}{TP+FN} = \frac{4}{4+3} = \frac{4}{7} = 57.14\%$$

F1:

$$F1: \frac{2TP}{2TP+FP+FN} = \frac{2(4)}{2(4)+2+3} = \frac{8}{13} = 61.54\%$$

1 c)

Given that IBM's HR wants to use this model to improve employee retention and identify, the goal is to identify all the employees who might have a high-risk of leaving and not just the accuracy of the model's prediction overall in which employee will be leaving and staying. In this case, the cost of misclassification for classifying an employee who will stay when they will actually leave is high since we want to improve employee retention; hence, recall would be a good measure to follow since high recall would mean the model is good at catching most of the employees at risk. Precision is less important than recall in this case because false alarms of checking in with an employee who is predicted to leave when they will actually stay is better than not checking in with an employee who is predicted to stay when they will actually leave.

2 1-NN, & Cross Validation(15 points)

2 a)

```
import pandas as pd
import numpy as np
data = {
    "ID": [1, 2, 3, 4, 5, 6, 7, 8, 9],
    "x1": [4.23, 2.15, 5.33, 3.49, 0.15, 8.23, 6.48, 1.53, 4.78],
    "x2": [7.01, 0.12, 5.14, 9.64, 4.23, 1.22, 2.09, 2.32, 3.67],
    "Class": ['-', '+', '-', '+', '-', '+', '-', '+', '-']
}

df = pd.DataFrame(data)

distance_matrix = np.zeros((df.shape[0], df.shape[0]))

def euclidean_distance(x1, x2):
    return np.sqrt(np.sum((x1 - x2) ** 2))

for i in range(df.shape[0]):
    for j in range(df.shape[0]):
        if i != j:
            distance_matrix[i, j] = euclidean_distance(df.iloc[i, 1:3], df.iloc[j, 1:3])

distance_matrix_df = pd.DataFrame(distance_matrix, columns=df["ID"])

distance_matrix_df
```

ID	1	2	3	4	5	6	7	8	9
ID									
1	0.000000	7.197117	2.169539	2.732124	4.937084	7.037336	5.410074	5.411663	3.384982
2	7.197117	0.000000	5.942457	9.613844	4.570788	6.178705	4.757079	2.285695	4.418077
3	2.169539	5.942457	0.000000	4.861646	5.259325	4.876105	3.259601	4.732061	1.569522
4	2.732124	9.613844	4.861646	0.000000	6.357964	9.662505	8.120505	7.577862	6.107782
5	4.937084	4.570788	5.259325	6.357964	0.000000	8.622442	6.681953	2.356374	4.663743
6	7.037336	6.178705	4.876105	9.662505	8.622442	0.000000	1.954329	6.789698	4.231430
7	5.410074	4.757079	3.259601	8.120505	6.681953	1.954329	0.000000	4.955341	2.320862
8	5.411663	2.285695	4.732061	7.577862	2.356374	6.789698	4.955341	0.000000	3.519233
9	3.384982	4.418077	1.569522	6.107782	4.663743	4.231430	2.320862	3.519233	0.000000

2 b)

i)

ID	Class	Predicted	Closest
1	-		
2	+		
3	-		
4	+		
5	-		
6	+	-	3
7	-	-	3
8	+	+	2
9	-	-	3

	Predicted Positive	**Predicted Negative**
Actual Positive	**1**	**1**
Actual Negative	**0**	**2**

Testing Accuracy: 75%

$$\frac{TP+TN}{TP+TN+FP+FN} = \frac{1+2}{1+2+0+1} = \frac{3}{4}$$

ii)

ID	Class	Predicted	Closest
1	-	+	4
2	+	+	8
3	-	-	9
4	+	-	1
5	-	+	8
6	+	-	7
7	-	+	6
8	+	+	2
9	-	-	3

	Predicted Positive	**Predicted Negative**
Actual Positive	**2**	**2**
Actual Negative	**3**	**2**

Testing Accuracy: 44.44%

$$\frac{TP+TN}{TP+TN+FP+FN} = \frac{2+2}{2+2+3+2} = \frac{4}{9}$$

iii)

ID	Class	Predicted	Closest
1	-	-	3
2	+	+	8
3	-	-	9
4	+	-	1
5	-	+	8
6	+	-	7
7	-	+	6
8	+	+	2
9	-	-	3

	Predicted Positive	**Predicted Negative**
Actual Positive	**2**	**2**
Actual Negative	**2**	**3**

Testing Accuracy: 55.56%

$$\frac{TP+TN}{TP+TN+FP+FN} = \frac{2+3}{2+2+3+2} = \frac{5}{9}$$

2 C)

Given the binary classification algorithm that uses majority vote classifier and three methods: holdout with 30/20 split, 5-fold cross validation, and LOOCV, it is likely the 0% validation accuracy comes from the LOOCV method. This is due to our majority vote classifier algorithm. Since there is a perfect balance between the positives and negatives, when an instance is left out and the model is trained with the rest of the data set, the training set will consist of 49 instances with the opposite binary sign from the one taken out being the majority in the training set. Hence, the validation accuracy will always be zero since the test set will always evaluate to the opposite binary sign due to the imbalance of the majority binary signs in the training set.

3 BN Inference(12 points)

Color	Car Type	Body Style	Popular
Black	Luxury	Sedan	Yes
Black	Sports	Sedan	No
Black	Sports	SUV	Yes
White	Luxury	Sedan	Yes
White	Luxury	SUV	No
White	Luxury	Sedan	No
Black	Luxury	SUV	Yes
Black	Luxury	Sedan	No
White	Sports	SUV	No
Black	Luxury	SUV	Yes

P(Popular = Yes)	5/10	P(Popular = No)	5/10
P(Color = Black | Popular = Yes)	4/5	P(Color = Black | Popular = No)	2/5
P(Color = White | Popular = Yes)	1/5	P(Color = White | Popular = No)	3/5
P(CarType = Luxury | Popular = Yes)	4/5	P(CarType = Luxury | Popular = No)	3/5
P(CarType = Sports | Popular = Yes)	1/5	P(CarType = Sports | Popular = No)	2/5
P(BodyStyle = Sedan | Popular = Yes)	2/5	P(BodyStyle = Sedan | Popular = No)	3/5
P(BodyStyle = SUV | Popular = Yes)	3/5	P(BodyStyle = SUV | Popular = No)	2/5

3 a)

{Color = Black, Car Type = Luxury, Body Style = Sedan}

Popular

3a) {Color = Black, Car Type = Luxury, Body Style = Sedan}

$P(\text{Color} = \text{Black}, \text{Car Type} = \text{Luxury}, \text{Body Style} = \text{Sedan} \mid \text{Popular} = \text{Yes})$

$4/5 \times 4/5 \times 2/5 \times 5/10 = 0.128$

$P(\text{Color} = \text{Black}, \text{Car Type} = \text{Luxury}, \text{Body Style} = \text{Sedan} \mid \text{Popular} = \text{No})$

$2/5 \times 3/5 \times 3/5 \times 5/10 = 0.072$

3 b)

{Color = Black, Car Type = Sports, Body Style = SUV }

Popular

3b) {Color = Black, Car Type = Sports, Body Style = SUV}

$P(\text{Color} = \text{Black}, \text{Car Type} = \text{Sports}, \text{Body Style} = \text{SUV} \mid \text{Popular} = \text{Yes})$

$= 4/5 \times 1/5 \times 3/5 \times 5/10 = 0.048$

$P(\text{Color} = \text{Black}, \text{Car Type} = \text{Sports}, \text{Body Style} = \text{SUV} \mid \text{Popular} = \text{No})$

$= 2/5 \times 2/5 \times 2/5 \times 5/10 = 0.032$

3 c)

{Color = White, Car Type = Sports, Body Style = Sedan}

Not Popular

3c) {Color = White, Car Type = Sports, Body Style = Sedan}

$P(\text{Color} = \text{White}, \text{CarType} = \text{Sports}, \text{Body Style} = \text{Sedan} \mid \text{Popular} = \text{Yes})$

$= 1/5 \times 1/5 \times 2/5 \times 5/10 = 0.008$

$P(\text{Color} = \text{White}, \text{CarType} = \text{Sports}, \text{Body Style} = \text{Sedan} \mid \text{Popular} = \text{No})$

$= 3/5 \times 2/5 \times 3/5 \times 5/10 = 0.072$

3 d)

{Color = White, Car Type = Luxury, Body Style = SUV }

Not Popular

3d) {Color = White, Car Type = Luxury, Body Style = SUV}

$P(\text{Color} = \text{White}, \text{CarType} = \text{Luxury}, \text{Body Style} = \text{SUV} \mid \text{Popular} = \text{Yes})$

$= 1/5 \times 4/5 \times 3/5 \times 5/10 = 0.048$

$P(\text{Color} = \text{White}, \text{CarType} = \text{Luxury}, \text{Body Style} = \text{SUV} \mid \text{Popular} = \text{No})$

$= 3/5 \times 3/5 \times 2/5 \times 5/10 = 0.072$