CS 513 MidTerm Report

Name: Shikhar Saxena
CWID: 20021187

1. A CART tree will be used to classify Blood Pressure using the features 'Gender' and 'Smoker'. Using the training data provided in the 'BP_cat_train' sheet of the 'BP_train_test' Excel file and 'BP_Status' (target):
   a. Utilize Excel to calculate the value of the Q(s/t) function for Gender=’Male’ vs. Gender=’Female’.
   b. Identify all possible first level splits of the classification and regression tree. (do not calculate the Q(s/t))
   (20 points)

Solution:

a. Make a pivot table to get the number of males and females with High and Normal Blood Pressure

PIVOT TABLE

Count of BP_Status	Column Labels					
Row Labels	**1- Non-smoker**	**2- Light**	**3- Moderate**	**4- Heavy**	**5- Very Heavy**	**Grand Total**
Female	**16**	**3**	**1**	**2**		**22**
High	8	1	1	1		11
Normal	8	2		1		11
Male	**5**		**1**	**8**	**4**	**18**
High	5			4	4	13
Normal			1	4		5
Grand Total	**21**	**3**	**2**	**10**	**4**	**40**

Goodness of a candidate split:

$$Q(s/t) = 2P_L * P_R \sum | p(j/t_L) - P(j/t_R) |$$

Here,

PL (Male): Count of Male instances / Total Count
PR (Male): Count of Female instances / Total Count
P(J/TL) (Male): Count of High BP in Male instances / Total Count of Male instances
P(J/TR) (Male): Count of High BP in Female instances / Total Count of Female instances
P(J/TL) (Female): Count of High BP in Female instances / Total Count of Female instances
P(J/TR) (Female): Count of High BP in Male instances / Total Count of Male instances

split	PL	PR	P(J/TL)		P(J/TR)		2PLPR	q(s/t)	Q(S/T)	
gender= male	18/40	22/40	13/18	5/18	11 /22	11/22	2*18*22/40*40 = .495	|13/18-11/22|+|5/18-11/22|= 0.4444	0.495*0.4444= .21997	
gender=female	22/40	18/40	11/22	11/22	5/18	13/18	2*18*22/40*40 = .495	|13/18-11/22|+|5/18-11/22|= 0.4444	0.495*0.4444= .21997	
Q(s/t) value is same for both Male and Female when target variable is BP_Status										

This is the calculation of each step

b. To identify all possible first-level splits for a classification and regression tree using the "Smoker" and "Gender" features, you would consider all combinations of possible splits. Since "Smoker" has 5 types (1- Non-smoker, 3- Moderate, 5- Very Heavy, 2- Light, 4- Heavy) and "Gender" has 2 (Male and Female), there are 5 + 2 = 7 possible first-level splits.

And they are as follows:

all possible first level splits of the classification and regression tree:

Candidate Splits for t = Root Node

Candidate	Left Child Node, tL	Right Child Node, tR
1	Somker= 1 Non-smoker	Somker =3- Moderate,5- Very Heavy,2- Light ,4-Heavy
2	Somker= 2 Light	Somker =1- Non-smoker, 3- Moderate,5- Very Heavy,4-Heavy
3	Gender = Female	Gender = Male
4	Somker= 3 Moderate	Somker =1- Non-smoker,5- Very Heavy,2- Light,4-Heavy
5	Somker= 4 Heavy	Somker =1- Non-smoker, 3- Moderate,5- Very Heavy,2- Light
6	Somker= 5 Very Heavy	Somker =1- Non-smoker, 3- Moderate ,2- Light,4-Heavy
7	Gender = Male	Gender = Female

2. Using the training data provided in the 'BP_cat_train' sheet of the 'BP_train_test' Excel file. Utilize Excel to build a categorical Naïve Bayes
model and score the ‘BP_cat_test’ using the features 'Gender' and 'Smoker'
as predictors and ‘BP_Status' as the target variable. (20 points)

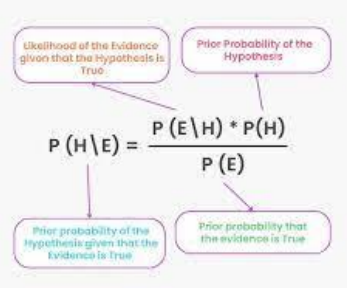
Solution:

Make a pivot table to get the number of males and females with High and Normal Blood Pressure using data given in ‘BP_cat_train’ sheet

PIVOT TABLE

Count of BP_Status	Column Labels					
Row Labels	**1- Non-smoker**	**2- Light**	**3- Moderate**	**4- Heavy**	**5- Very Heavy**	**Grand Total**
Female	**16**	**3**	**1**	**2**		**22**
High	8	1	1	1		11
Normal	8	2		1		11
Male	**5**		**1**	**8**	**4**	**18**
High	5			4	4	13
Normal			1	4		5
Grand Total	**21**	**3**	**2**	**10**	**4**	**40**

Calculate the conditional probabilities using the data obtained in the pivot table



Conditional Probability formula

			Probablity Of Gender given High BP			Probablity of Smoker Given High BP				
Total People with High BP	13+11= 24		Male	Female		Non-Smoker	Light	Moderate	Heavy	Very Heavy
Total People with Normal BP	5+11=16		13/24 = 0.541	11/24 = 0.458		13/24=0.541	1/24=0.041	1/24=0.041	5/24=0.208	4/24=0.166
Total Male with High BP	13			Probablity of High BP	24/40=0.6					
Total Male with Normal BP	5			Probablity of Normal BP	16/40=0.4					
Total Female with High BP	11		Probablity Of Gender given Normal BP			Probablity of Smoker Given normal BP				
Total Female with Normal BP	11		Male	Female		Non-Smoker	Light	Moderate	Heavy	Very Heavy
			5/16=0.312	11/16=0.687		8/16=0.5	2/16=0.125	1/16=0.062	5/16=0.312	0

BP_cat_test:

Smoker	Gender	BP_Status
2- Light	Female	High
5- Very Heavy	Female	High
3- Moderate	Female	Normal
4- Heavy	Male	Normal
1- Non-smoker	Female	Normal

Calculate the probabilities for High BP and Normal BP for all the five testing data and declare the prediction as the one whichever has a higher value.

The table shows the results after calculations

	Light & Female	Very Heavy & Female	Moderrate & Female	Heavy &Male	Non-smoker & Female
Probablity of High BP	(0.041*0.458*0.6)/0.075*0.55=**0.273**	(0.166*0.458*0.6)/0.1*0.55=**0.829**	0.041*0.458*0.6/0.05*0.55=**0.409**	0.208*0.541*0.6/0.25*0.45=**0.6**	0.541*0.458*0.6/0.525*0.55=**0.514**
Probablity OF Normal BP	(0.125*0.687*0.4)/0.075*0.55=**0.832**	0	0.062*0.687*0.4/0.05*0.55=**0.613**	0.312*0.312*0.4/0.25*0.45=**0.346**	0.5*0.687*0.4/0.525*0.55=**0.475**
Predictions	**Normal BP**	**High BP**	**Normal BP**	**High BP**	**High BP**

Calculate the metrics :

CWID			20021187
Name		Shikhar Saxena	
Date			08-11-2023

Confusion Matrix			
		Predicted	
		Normal	High
Actual	Normal	1	2
	High	1	1

Results	
Accuracy	2/5=0.4
Precision	1/3=0.3333
F1	1/2=0.5
Recall	2/5=0.4

Here,

Accuracy = (Number of True Positives + Number of True Negatives) / (Total Number of Predictions)
Precision = Number of True Positives / (Number of True Positives + Number of False Positives)
Recall = Number of True Positives / (Number of True Positives + Number of False Negatives)
F1 score = 2 * (Precision * Recall) / (Precision + Recall)

3. Using the training data provided in the 'BP-Num_train' sheet of the 'BP_train_test' Excel file, use Excel to build a knn model and score the 'BP_Num_test' using the features 'Gender', 'Age', 'height' and 'weight' as
predictors and 'BP_Status' as the target variable. Use: K=2, weighted, Euclidian distance. (20 points)

	Min	Max
Age (years)	25	65
Weight (lb)	50	300
Height (inch)	50	80

Solution:

here's the normalised (MinMax normalisation)training and testing data:

Testing Data	Normalized Values			
Gender	Age	Height	Weight	BP_Status
Female	0.45	0.41667	0.656	High
Female	0.5	0.33333	0.424	High
Female	0.35	0.35833	0.348	Normal
Male	0.875	0.56667	0.444	Normal
Female	0.6	0.21667	0.32	Normal
Training Data				
Gender	**Age**	**Height**	**Weight**	**BP_Status**
Female	0.8	0.40833	0.328	High
Female	0.575	0.31667	0.348	Normal
Male	0.5	0.4	0.4	High
Female	0.25	0.41667	0.228	Normal
Male	0.575	0.48333	0.416	High
Female	0.4	0.43333	0.416	Normal
Male	0.25	0.58333	0.468	High
Female	0.8	0.46667	0.36	High
Female	0.55	0.38333	0.216	Normal
Male	0.3	0.61667	0.404	Normal
Female	0.725	0.45	0.404	High
Male	0.175	0.65	0.484	Normal
Female	0.5	0.36667	0.32	Normal
Female	0.65	0.475	0.528	High
Female	0.8	0.36667	0.44	High
Male	0.55	0.575	0.412	Normal
Female	0.675	0.30833	0.296	Normal
Male	0.5	0.54167	0.568	High
Male	0.325	0.59167	0.432	High
Female	0.475	0.48333	0.312	Normal
Female	0.25	0.43333	0.232	Normal
Male	0.325	0.55833	0.4	High
Male	0.575	0.48333	0.548	High
Male	0.5	0.65833	0.548	Normal
Female	0.575	0.65833	0.592	High
Female	0.8	0.40833	0.34	High
Male	0.375	0.55	0.484	High
Male	0.875	0.475	0.688	High
Female	0.15	0.3	0.304	Normal
Female	0.7	0.54167	0.468	High
Female	0.15	0.375	0.312	Normal
Male	0.425	0.45833	0.268	High
Female	0.6	0.525	0.368	High
Male	0.325	0.60833	0.484	High
Female	0.775	0.34167	0.416	Normal
Male	0.225	0.45	0.28	Normal
Male	0.55	0.45	0.54	High
Male	0.525	0.625	0.496	High
Female	0.225	0.39167	0.368	High
Female	0.275	0.41667	0.264	High

Next, calculate the Euclidian distance of all the five data points in testing data from the training data using the given formula:

$$d(x,\ y)\ = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

For excel, this formula would be used:

Formula for calculating Euclidean Distance

```
=SQRT((B10-$B$3)^2 + (C10-$C$3)^2 + (D10-$D$3)^2)
```

Test 1	Test 2	Test 3	Test 4	Test 5
Euclidean Distance	**Euclidean Distance**	**Euclidean Distance**	**Euclidean Distance**	**Euclidean Distance**
0.479743103	0.3237916	0.453210768	0.210120071	0.27712833
0.347115255	0.108068394	0.228825504	0.402139279	0.106812921
0.261369045	0.070855095	0.16413443	0.412721186	0.223631642
0.472423539	0.328421139	0.16674165	0.67807153	0.413477932
0.278692383	0.167895801	0.266221712	0.31261549	0.284520845
0.2457189	0.14164745	0.11291147	0.494152586	0.310097476
0.321125797	0.356280788	0.273906918	0.625682649	0.528061023
0.461103025	0.334475377	0.463011999	0.150602125	0.322645316
0.45245012	0.219690692	0.240933601	0.437287218	0.202716003
0.354970421	0.347387072	0.269020652	0.578554233	0.507006903
0.374486463	0.254236329	0.390081758	0.194193489	0.277714682
0.399566571	0.457714734	0.366320139	0.706076798	0.628727904
0.34335987	0.109211314	0.152818338	0.442720002	0.180277564
0.244513349	0.231052904	0.368796843	0.257067652	0.335410362
0.414313891	0.302269931	0.459383766	0.213637544	0.277308492
0.307579981	0.247076461	0.301729091	0.326677891	0.373318601
0.438133668	0.218252148	0.332909898	0.358664343	0.120846091
0.160838428	0.253256348	0.323281783	0.395760028	0.420866962
0.310525361	0.312129959	0.249249763	0.550698647	0.478324158
0.351291111	0.188862384	0.1804051	0.429381467	0.294618586
0.4690989	0.330702283	0.170531522	0.673310313	0.420937578
0.318167321	0.286052443	0.208156191	0.55182012	0.445826324
0.178138835	0.208568933	0.32596012	0.32826886	0.351738697
0.269382215	0.34785198	0.390512484	0.399804675	0.507004383
0.279506311	0.37346218	0.447393563	0.346852674	0.519305733
0.471620021	0.320438762	0.452839928	0.203741121	0.277733885
0.23018857	0.257234221	0.236341091	0.501874265	0.434317984
0.43017645	0.479990046	0.636267327	0.260650682	0.527053234
0.476985441	0.371498467	0.212929044	0.785070768	0.457930611
0.336851599	0.292128016	0.412929911	0.17839843	0.370849026
0.458336242	0.369837953	0.203896488	0.761436216	0.477109468
0.391030831	0.213508782	0.148408221	0.495188965	0.302872544
0.342315806	0.223320646	0.301127511	0.288335067	0.312047183
0.286260565	0.331436268	0.285693892	0.553024512	0.505889096
0.410913616	0.275242519	0.430728195	0.247808394	0.235512208
0.439445231	0.331620432	0.16926836	0.680446259	0.443474288
0.15673899	0.171950897	0.292004756	0.358402164	0.324568089
0.273180852	0.301460519	0.351624958	0.358617872	0.450929164
0.366324992	0.286642247	0.13090497	0.67742232	0.416598128
0.429288947	0.288391131	0.126821835	0.644127317	#REF!

For k=1 and k=2, calculate the weighted distance and predict BP_Status for all the 5 testing values.

	K=2		Weighted Distance		BP_status		Predicted BP_status	Pass/Fail
	1	2						
1	0.15673899	0.160838428	6.380033439	6.217419624	High	Normal	High	Pass
2	0.070855095	0.108068394	14.11331118	9.253399289	High	High	High	Pass
3	0.11291147	0.126821835	8.856496176	7.885077532	Normal	High	Normal	Pass
4	0.150602125	0.17839843	6.640012558	5.605430481	High	Normal	High	Fail
5	#REF!	#REF!	#REF!	#REF!	#REF!	#REF!	#REF!	Pass

Formula for calculating KNN
=SMALL($G$3:$G$42, 1)

Formula for getting the status
=INDEX($E$10:$E$49, MATCH(N43, G3:G42, 0))

Here:

The formula =SMALL($G$3:$G$42, 1) in Excel is used to find the smallest (or in this case, the 1st smallest) value within the specified range $G$3:$G$42.

Use the INDEX function to select the top k rows

Calculate the metrics:

Confusion Matrix

		Predicted	
		Normal	High
Actual	Normal	2	1
	High	0	2

Accuracy = (TP + TN) / (TP + TN + FP + FN)
=(2 + 2) / (2 + 2 + 1 + 0)
=4 / 5
=0.8 (or 80%)

F1 Score = 2 * (Precision * Recall) /
(Precision + Recall)
= 2 * (0.6667 * 1) / (0.6667 + 1)
= 1.0

Precision = TP / (TP + FP)
= 2 / (2 + 1)
= 2 / 3
≈ 0.6667 (or 66.67%)