Programming Assignment # 2 Classifiers

The purpose of this assignment is to derive different classifiers. This assignment is to classify drug users using Decision Tree, Bayes Classifier, nearest neighbor and to use classification metrics.

Rules

1. Work is to be done individually
2. Any cheating including plagiarism, collusion will be reported to the corresponding UTA' s instance
3. If using any resource (books, internet) , please make sure that you cite it.

- Use Panda, Jupyter notebook, scikit for the following assignment.
- For the Decision Tree use entropy as a classifier criteria
- For Nearest neighbor use k = 3 and Euclidean as the distance
- For Bayes, consider Laplace smoothing and gaussian distribution considering that the attributes are continuous.
- For classification metrics (http://scikit-learn.org/stable/modules/model_evaluation.html#classification-metrics)

Instructions:

Use 70 % of the dataset for training and 30 % for Testing.

Place all your tasks in ONE NOTEBOOK BUT SEPARATE BLOCKS folder named lastname_prog2. Compress your folder and submit it.

Tasks:

Task # 1: Transform data into a binary classification by union of part of classes into one new class. There will be two classes **NOUSER**, **USER**. Use "Never Used", "Used over a Decade Ago" , Used in Last decade to form class **NOUSER** , and all other classes form class "**USER**". 5 points
Use the new data set for Task 2,3,4,5

Task # 2; Define a decision tree using Entropy as the criteria. Show your tree (use Graphviz). 20 points

Task # 3 Define a classifier using Bayes classifier, take into consideration Laplace smoothing. Use gaussian Naïve Bayes if necessary, depending on the attribute (continuous or discrete). 20 points

Task # 4: Use classification metrics to display precision, confusion matrix, classification report for each one of the trees. 30 points

Task # 5; Define a classifier using KNN 25 points
The data set is given as text file and a csv file.

Task # 6:

Bonus question (10 points) :

Use correlation and Random Forest for Feature selection.

Data Set Information:

Database contains records for 1885 respondents. For each respondent 12 attributes are known: 1) **Personality measurements** which include NEO-FFI-R (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness), BIS-11 (impulsivity), and ImpSS (sensation seeking), 2) **level of education**, 3) **age,** 4) **gender,** 5) **country of residence and 6) ethnicity**. All input attributes are originally categorical and are quantified. After quantification values of all input features can be considered as real-valued. **In addition, participants were questioned concerning their use of 14 legal and illegal drugs** (amphetamines, amyl nitrite, benzodiazepine, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, and volatile substance abuse and one fictitious drug (Semeron) which was introduced to identify over-claimers. For each drug they have to select one of the answers: never used the drug, used it over a decade ago, or in the last decade, year, month, week, or day.
Database contains 14 classification problems. Each of independent label variables contains seven classes: "Never Used", "Used over a Decade Ago", "Used in Last Decade", "Used in Last Year", "Used in Last Month", "Used in Last Week", and "Used in Last Day".

Attribute Information:

a. ID is number of record in original database. Cannot be related to participant. It can be used for reference only.

b. Age (Real) is age of participant and has one of the values:
Value Meaning Cases Fraction
-0.95197 18-24 643 34.11%
-0.07854 25-34 481 25.52%
0.49788 35-44 356 18.89%
1.09449 45-54 294 15.60%
1.82213 55-64 93 4.93%
2.59171 65+ 18 0.95%
Descriptive statistics
Min Max Mean Std.dev.
-0.95197 2.59171 0.03461 0.87813

c. Gender (Real) is gender of participant:
Value Meaning Cases Fraction
0.48246 Female 942 49.97%
-0.48246 Male 943 50.03%
Descriptive statistics
Min Max Mean Std.dev.
-0.48246 0.48246 -0.00026 0.48246

d. Education (Real) is level of education of participant and has one of the values:
Value Meaning Cases Fraction
-2.43591 Left school before 16 years 28 1.49%

-1.73790 Left school at 16 years 99 5.25%
-1.43719 Left school at 17 years 30 1.59%
-1.22751 Left school at 18 years 100 5.31%
-0.61113 Some college or university, no certificate or degree 506 26.84%
-0.05921 Professional certificate/ diploma 270 14.32%
0.45468 University degree 480 25.46%
1.16365 Masters degree 283 15.01%
1.98437 Doctorate degree 89 4.72%
Descriptive statistics
Min Max Mean Std.dev.
-2.43591 1.98437 -0.00379 0.95004

e. Country (Real) is country of current residence of participant and has one of the values:
Value Meaning Cases Fraction
-0.09765 Australia 54 2.86%
0.24923 Canada 87 4.62%
-0.46841 New Zealand 5 0.27%
-0.28519 Other 118 6.26%
0.21128 Republic of Ireland 20 1.06%
0.96082 UK 1044 55.38%
-0.57009 USA 557 29.55%
Descriptive statistics
Min Max Mean Std.dev.
-0.57009 0.96082 0.35554 0.70015

f. Ethnicity (Real) is ethnicity of participant and has one of the values:
Value Meaning Cases Fraction
-0.50212 Asian 26 1.38%
-1.10702 Black 33 1.75%
1.90725 Mixed-Black/Asian 3 0.16%
0.12600 Mixed-White/Asian 20 1.06%
-0.22166 Mixed-White/Black 20 1.06%
0.11440 Other 63 3.34%
-0.31685 White 1720 91.25%
Descriptive statistics
Min Max Mean Std.dev.
-1.10702 1.90725 -0.30958 0.16618

g. Nscore (Real) is NEO-FFI-R Neuroticism. Possible values are presented in table below:
Nscore Cases Value Nscore Cases Value Nscore Cases Value
12 1 -3.46436 29 60 -0.67825 46 67 1.02119
13 1 -3.15735 30 61 -0.58016 47 27 1.13281
14 7 -2.75696 31 87 -0.46725 48 49 1.23461
15 4 -2.52197 32 78 -0.34799 49 40 1.37297
16 3 -2.42317 33 68 -0.24649 50 24 1.49158
17 4 -2.34360 34 76 -0.14882 51 27 1.60383
18 10 -2.21844 35 69 -0.05188 52 17 1.72012
19 16 -2.05048 36 73 0.04257 53 20 1.83990
20 24 -1.86962 37 67 0.13606 54 15 1.98437
21 31 -1.69163 38 63 0.22393 55 11 2.12700
22 26 -1.55078 39 66 0.31287 56 10 2.28554
23 29 -1.43907 40 80 0.41667 57 6 2.46262
24 35 -1.32828 41 61 0.52135 58 3 2.61139
25 56 -1.19430 42 77 0.62967 59 5 2.82196
26 57 -1.05308 43 49 0.73545 60 2 3.27393
27 65 -0.92104 44 51 0.82562

28 70 -0.79151 45 37 0.91093
Descriptive statistics
Min Max Mean Std.dev.
-3.46436 3.27393 0.00004 0.99808

h. Escore (Real) is NEO-FFI-R Extraversion. Possible values are presented in table below:
Escore Cases Value Escore Cases Value Escore Cases Value
16 2 -3.27393 31 55 -1.23177 45 91 0.80523
18 1 -3.00537 32 52 -1.09207 46 69 0.96248
19 6 -2.72827 33 77 -0.94779 47 64 1.11406
20 3 -2.53830 34 68 -0.80615 48 62 1.28610
21 3 -2.44904 35 58 -0.69509 49 37 1.45421
22 8 -2.32338 36 89 -0.57545 50 25 1.58487
23 5 -2.21069 37 90 -0.43999 51 34 1.74091
24 9 -2.11437 38 106 -0.30033 52 21 1.93886
25 4 -2.03972 39 107 -0.15487 53 15 2.12700
26 21 -1.92173 40 130 0.00332 54 10 2.32338
27 23 -1.76250 41 116 0.16767 55 9 2.57309
28 23 -1.63340 42 109 0.32197 56 2 2.85950
29 32 -1.50796 43 105 0.47617 58 1 3.00537
30 38 -1.37639 44 103 0.63779 59 2 3.27393
Descriptive statistics
Min Max Mean Std.dev.
-3.27393 3.27393 -0.00016 0.99745

i. Oscore (Real) is NEO-FFI-R Openness to experience. Possible values are presented in table below:
Oscore Cases Value Oscore Cases Value Oscore Cases Value
24 2 -3.27393 38 64 -1.11902 50 83 0.58331
26 4 -2.85950 39 60 -0.97631 51 87 0.72330
28 4 -2.63199 40 68 -0.84732 52 87 0.88309
29 11 -2.39883 41 76 -0.71727 53 81 1.06238
30 9 -2.21069 42 87 -0.58331 54 57 1.24033
31 9 -2.09015 43 86 -0.45174 55 63 1.43533
32 13 -1.97495 44 101 -0.31776 56 38 1.65653
33 23 -1.82919 45 103 -0.17779 57 34 1.88511
34 25 -1.68062 46 134 -0.01928 58 19 2.15324
35 26 -1.55521 47 107 0.14143 59 13 2.44904
36 39 -1.42424 48 116 0.29338 60 7 2.90161
37 51 -1.27553 49 98 0.44585
Descriptive statistics
Min Max Mean Std.dev.
-3.27393 2.90161 -0.00053 0.99623

j. Ascore (Real) is NEO-FFI-R Agreeableness. Possible values are presented in table below:
Ascore Cases Value Ascore Cases Value Ascore Cases Value
12 1 -3.46436 34 42 -1.34289 48 104 0.76096
16 1 -3.15735 35 45 -1.21213 49 85 0.94156
18 1 -3.00537 36 62 -1.07533 50 68 1.11406
23 1 -2.90161 37 83 -0.91699 51 58 1.2861
24 2 -2.78793 38 82 -0.76096 52 39 1.45039
25 1 -2.70172 39 102 -0.60633 53 36 1.61108
26 7 -2.53830 40 98 -0.45321 54 36 1.81866
27 7 -2.35413 41 114 -0.30172 55 16 2.03972
28 8 -2.21844 42 101 -0.15487 56 14 2.23427
29 13 -2.07848 43 105 -0.01729 57 8 2.46262
30 18 -1.92595 44 118 0.13136 58 7 2.75696

31 24 -1.77200 45 112 0.28783 59 1 3.15735
32 30 -1.62090 46 100 0.43852 60 1 3.46436
33 34 -1.47955 47 100 0.59042
Descriptive statistics
Min Max Mean Std.dev.
-3.46436 3.46436 -0.00024 0.99744

k. Cscore (Real) is NEO-FFI-R Conscientiousness. Possible values are presented in table below:
Cscore Cases Value Cscore Cases Value Cscore Cases Value
17 1 -3.46436 32 39 -1.25773 46 113 0.58489
19 1 -3.15735 33 49 -1.13788 47 95 0.7583
20 3 -2.90161 34 55 -1.01450 48 95 0.93949
21 2 -2.72827 35 55 -0.89891 49 76 1.13407
22 5 -2.57309 36 69 -0.78155 50 47 1.30612
23 5 -2.42317 37 81 -0.65253 51 43 1.46191
24 6 -2.30408 38 77 -0.52745 52 34 1.63088
25 9 -2.18109 39 87 -0.40581 53 28 1.81175
26 13 -2.04506 40 97 -0.27607 54 27 2.04506
27 13 -1.92173 41 99 -0.14277 55 13 2.33337
28 25 -1.78169 42 105 -0.00665 56 8 2.63199
29 24 -1.64101 43 90 0.12331 57 3 3.00537
30 29 -1.51840 44 111 0.25953 59 1 3.46436
31 41 -1.38502 45 111 0.41594
Descriptive statistics
Min Max Mean Std.dev.
-3.46436 3.46436 -0.00039 0.99752

l. Impulsive (Real) is impulsiveness measured by BIS-11. Possible values are presented in table below:
Impulsiveness Cases Fraction
-2.55524 20 1.06%
-1.37983 276 14.64%
-0.71126 307 16.29%
-0.21712 355 18.83%
0.19268 257 13.63%
0.52975 216 11.46%
0.88113 195 10.34%
1.29221 148 7.85%
1.86203 104 5.52%
2.90161 7 0.37%
Descriptive statistics
Min Max Mean Std.dev.
-2.55524 2.90161 0.00721 0.95446

m. SS (Real) is sensation seeing measured by ImpSS. Possible values are presented in table below:
SS Cases Fraction
-2.07848 71 3.77%
-1.54858 87 4.62%
-1.18084 132 7.00%
-0.84637 169 8.97%
-0.52593 211 11.19%
-0.21575 223 11.83%
0.07987 219 11.62%
0.40148 249 13.21%
0.76540 211 11.19%
1.22470 210 11.14%
1.92173 103 5.46%

Descriptive statistics
Min Max Mean Std.dev.
-2.07848 1.92173 -0.00329 0.96370

n. Amphet is class of amphetamines consumption. It is output attribute with following distribution of classes.

o. Amyl is class of amyl nitrite consumption. It is output attribute with following distribution of classes.

p. Benzos is class of benzodiazepine consumption. It is output attribute with following distribution of classes:

Value Class	Alcohol		Amphet		Amyl		Benzos	
	Cases	Fraction	Cases	Fraction	Cases	Fraction	Cases	Fraction
CL0 Never Used	34	1.80%	976	51.78%	1305	69.23%	1000	53.05%
CL1 Used over a Decade Ago	34	1.80%	230	12.20%	210	11.14%	116	6.15%
CL2 Used in Last Decade	68	3.61%	243	12.89%	237	12.57%	234	12.41%
CL3 Used in Last Year	198	10.50%	198	10.50%	92	4.88%	236	12.52%
CL4 Used in Last Month	287	15.23%	75	3.98%	24	1.27%	120	6.37%
CL5 Used in Last Week	759	40.27%	61	3.24%	14	0.74%	84	4.46%
CL6 Used in Last Day	505	26.79%	102	5.41%	3	0.16%	95	5.04%

q. Coke is class of cocaine consumption. It is output attribute with following distribution of classes:

Value Class	Caff		Cannabis		Choc		Coke	
	Cases	Fraction	Cases	Fraction	Cases	Fraction	Cases	Fraction
CL0 Never Used	27	1.43%	413	21.91%	32	1.70%	1038	55.07%
CL1 Used over a Decade Ago	10	0.53%	207	10.98%	3	0.16%	160	8.49%
CL2 Used in Last Decade	24	1.27%	266	14.11%	10	0.53%	270	14.32%
CL3 Used in Last Year	60	3.18%	211	11.19%	54	2.86%	258	13.69%
CL4 Used in Last Month	106	5.62%	140	7.43%	296	15.70%	99	5.25%
CL5 Used in Last Week	273	14.48%	185	9.81%	683	36.23%	41	2.18%
CL6 Used in Last Day	1385	73.47%	463	24.56%	807	42.81%	19	1.01%

r. Crack is class of crack consumption. It is output attribute with following distribution of classes.

s. Ecstasy is class of ecstasy consumption. It is output attribute with following distribution of classes.

t. Heroin is class of heroin consumption. It is output attribute with following distribution of classes.

u. Ketamine is class of ketamine consumption. It is output attribute with following distribution of classes:

Value Class	Crack		Ecstasy		Heroin		Ketamine	
	Cases	Fraction	Cases	Fraction	Cases	Fraction	Cases	Fraction
CL0 Never Used	1627	86.31%	1021	54.16%	1605	85.15%	1490	79.05%
CL1 Used over a Decade Ago	67	3.55%	113	5.99%	68	3.61%	45	2.39%
CL2 Used in Last Decade	112	5.94%	234	12.41%	94	4.99%	142	7.53%
CL3 Used in Last Year	59	3.13%	277	14.69%	65	3.45%	129	6.84%
CL4 Used in Last Month	9	0.48%	156	8.28%	24	1.27%	42	2.23%
CL5 Used in Last Week	9	0.48%	63	3.34%	16	0.85%	33	1.75%
CL6 Used in Last Day	2	0.11%	21	1.11%	13	0.69%	4	0.21%

v. Legalh is class of legal highs consumption. It is output attribute with following distribution of classes

w. LSD is class of alcohol consumption. It is output attribute with following distribution of classes

x. Meth is class of methadone consumption. It is output attribute with following distribution of classes.

y. Mushrooms is class of magic mushrooms consumption. It is output attribute with following distribution

of classes:
Value Class Legalh LSD Meth Mushrooms
Cases Fraction Cases Fraction Cases Fraction Cases Fraction
CL0 Never Used 1094 58.04% 1069 56.71% 1429 75.81% 982 52.10%
CL1 Used over a Decade Ago 29 1.54% 259 13.74% 39 2.07% 209 11.09%
CL2 Used in Last Decade 198 10.50% 177 9.39% 97 5.15% 260 13.79%
CL3 Used in Last Year 323 17.14% 214 11.35% 149 7.90% 275 14.59%
CL4 Used in Last Month 110 5.84% 97 5.15% 50 2.65% 115 6.10%
CL5 Used in Last Week 64 3.40% 56 2.97% 48 2.55% 40 2.12%
CL6 Used in Last Day 67 3.55% 13 0.69% 73 3.87% 4 0.21%

z. Semer is class of fictitious drug Semeron consumption. It is output attribute with following distribution of classes.

aa. VSA is class of volatile substance abuse consumption. It is output attribute with following distribution of classes:
Value Class Semer VSA
Cases Fraction Cases Fraction Cases Fraction
CL0 Never Used 428 22.71% 1877 99.58% 1455 77.19%
CL1 Used over a Decade Ago 193 10.24% 2 0.11% 200 10.61%
CL2 Used in Last Decade 204 10.82% 3 0.16% 135 7.16%
CL3 Used in Last Year 185 9.81% 2 0.11% 61 3.24%
CL4 Used in Last Month 108 5.73% 1 0.05% 13 0.69%
CL5 Used in Last Week 157 8.33% 0 0.00% 14 0.74%
CL6 Used in Last Day 610 32.36% 0 0.00% 7 0.37%